The Expression of a Need

Understanding Search

Alfort, Esben

Document Version
Final published version

Publication date:
2013

COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN
SOLBJERG PLADS 3
DK-2000 FREDERIKSBERG
DANMARK

www.cbs.dk

ISSN 0906-6934

Print ISBN: **978-87-92977-60-1**
Online ISBN: **978-87-92977-61-8**

The Expression of a Need

PhD Series 24-2013

Copenhagen Business School
HANDELSHØJSKOLEN

The Expression of a Need

Understanding search

Esben Alfort

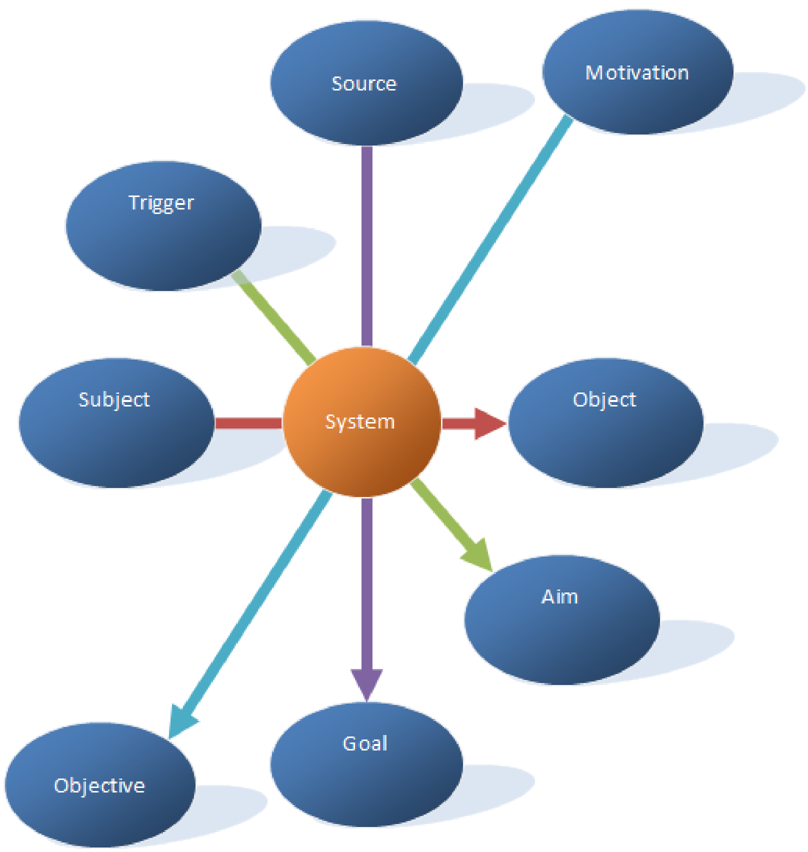
LIMAC PhD School
Department of International Business Communication

PhD Series 24-2013

The Expression of a Need

- Understanding search

Esben Alfort



LIMAC PhD School

Copenhagen Business School

Esben Alfort
The Expression of a Need
Understanding search

1st edition 2013
PhD Series 24.2013

ISSN 0906-6934

Print ISBN: 978-87-92977-60-1
Online ISBN: 978-87-92977-61-8

LIMAC PhD School is a cross disciplinary PhD School connected to research communities within the areas of Languages, Law, Informatics, Operations Management, Accounting, Communication and Cultural Studies.

The expression of a need – Understanding search

Esben Alfort

Ankiro / Copenhagen Business School

esben@ankiro.dk

Abstract

This thesis provides a framework for information retrieval based on a set of models which together illustrate how users of search engines come to express their needs in a particular way. With such insights, we may be able to improve systems' capabilities of understanding users' requests and through that eventually the ability to satisfy their needs. Developing the framework necessitates discussion of context, relevance, need development, and the cybernetics of search, all of which are controversial topics. Transaction log data from two enterprise search engines are analysed using a specially developed method which classifies queries according to what aspect of the need they refer to.

Denne afhandling bidrager med et framework for informationssøgning baseret på en række modeller der tilsammen illustrerer hvordan brugere af en søgemaskine når frem til at formulere deres behov på en bestemt måde. Dette tillader os at forbedre systemernes mulighed for at forstå brugernes forespørgsler og dermed på sigt hjælpe dem med at få opfyldt deres behov i højere grad. Udviklingen af et sådant framework har nødvendiggjort behandling af en række kontroversielle emner, såsom kontekst, relevans, behovsudvikling og søgeprocessens cybernetiske natur. Logdata fra enterprise search på to forskellige websites analyseres ved hjælp af en særligt udviklet metode, som klassificerer søgestrenge efter hvilket aspekt ved brugerens behov de refererer til.

Contents

Preface

> *The importance for progress in IR research is to have an idea of when to use the different principles and techniques, to have precise understanding of parameters for their appropriate application. (Ingwersen, 1992)*
>
> *Our intention in bringing together these diverse elements is to generate understanding – to let the apparent chasms become spaces in which new possibilities are revealed. (Winograd & Flores, 1986)*

I have always made it my business to make an effort to really understand the fundamental processes behind the phenomena I settle down to investigate – the ones that are so basic that they are often left alone and forgotten, and yet everything hinges on the correct understanding of them. Right from my humble beginnings in the linguistics department of Copenhagen University some thirteen years ago I realized that the deeper you dig into current linguistic theories the more assumptions they seem to rest on, and these often grow more and more tenuous and questionable along the way. It soon became my mission to uncover the roots of such theories and unravel the fundamental assumptions behind them in the hope of being able to decide whether these were likely to be correct, and possibly to correct some obsolete views that by a trick of history had come to support whole complexes of advanced theories formulated by later, often brilliant, researchers who had just failed to look back in sufficient detail. This thesis is another brick in that wall of understanding. The field of **IR research** is currently so fragmented and governed by so many divergent interests that few people settle down and try to assemble and connect findings and results into a coherent picture that will actually make us *understand* what is going on in the interaction between humans with needs and systems designed to help them satisfy those needs, though as we shall see there are some notable exceptions[1]. A related field that is hampered by the sheer abstractness of its subject is **relevance research**, resulting in a situation in which every new researcher claims to have identified *the* interesting feature of relevance and thus introduces a new type. Focusing so narrowly on one type does not contribute much to our overall understanding of the phenomenon, though it may provide vital data on one particular aspect of it. This

[1] One particularly good example is Ingwersen and Järvelin's (2005a) book.

thesis attempts to zoom out and look at the very human phenomena of relevance assessment, information needs and search strategies from a wider and more explanatory perspective.

The present work should be seen in the light of the overarching mission to uncover the basic building blocks of information retrieval and of language as used in search, and more particularly in enterprise search as opposed to web search. It is a transdisciplinary study in the sense that it attempts to combine insights from linguistics, cognitive science, philosophy, cybernetics, library and information studies, lexicography, information retrieval research, and context-awareness and adaptability studies. Transdisciplinarity inevitably comes at a price to the individual researcher. Firstly, it is inconceivable for one researcher to be exhaustively familiar with the accumulating literatures belonging to all of these branches, and secondly, the risk of using terms that have different meanings or connotations in the various fields is unavoidable. The reader is entreated to bear these difficulties in mind whenever an account seems to be particularly incomplete from the point of view of one field, or a notion seems to be given a definition that does not agree with the accepted terminology of a certain research tradition. I have tried to be explicit in my definitions, but if in doubt, the intended meaning of a term is more likely to be one drawn from linguistics than from any other branch, due to my own roots in that field. It is certainly true that it has not been possible in the time I have been working on this project to cover anything like the entire literature, which is truly vast, given the many different fields involved.

The thesis opens up a field of investigation that has to my knowledge never been studied in depth before, viz. the detailed linguistic analysis of queries and of what people choose to refer to, in relation to what they hope to achieve. No in-depth description of these matters can be given at this point, but at least this work presents a beginning on which to build in future. Understanding the more or less conscious choices that people make when expressing their needs in the form of a query necessitates a model of what they have to choose from. It is the construction of such a model that is the main achievement of this thesis. On the way to this goal it has been deemed necessary to dig deep into longstanding discussions of the workings of highly elusive phenomena such as relevance, need, intention and context. An attempt is made below at building a thorough understanding of them, too, lest we advance on insecure ground.

One of these concepts is particularly unavoidable: The delightfully enigmatic phenomenon of *relevance* is at the centre of all information retrieval research and has never ceased to attract heated debate and ingenious interpretations. Understanding what makes a resource relevant is crucial if we are to be able to provide the right material to those who need it. However, relevance has proven a particularly elusive phenomenon that continues to evade definition, and still has not been properly understood. This seems largely due to the immense complexity of the phenomenon, which depends on a range of other phenomena that are equally hard to typologize, such as user needs and contextual parameters.

Typologizing without a clear understanding of what criteria to base the distinction on merely makes matters worse. Unfortunately, this is a common state of affairs in relevance studies and information retrieval research in general, where it can be really hard to grasp what belongs where. In this thesis, great emphasis is consequently laid on producing consistent typologies and models, both as regards user needs and goal types, contextual parameters, and relevance criteria. It is hoped that the effort put into this will help make matters clearer to future researchers.

Nature of the contributions

The aim of this thesis is to develop a framework consisting of a set of models that together explain query variation and through that enable better query log analysis. The end product is the Faceted Stimulus model (Part I), the Regulated Flux model (Part II) and the Stage-based Intent model (Part III), which together describe the search process and explain the choices that users have when deciding how to phrase their queries, as well as a Facet-based Context model (Part I) and a method of analysing search logs called Target Analysis (Part III). The models do not necessarily predict how users will actually choose to express themselves, but they help understand the issues involved and allow future research into the matter. As mentioned above, the main goal of this thesis is the general one of promoting a better understanding of the issues involved rather than suggesting exact solutions. Frequently, simply combining insights from different researchers and fields into a coherent whole is enough, because the individual researchers may have had brilliant suggestions, yet have focused on too specific cases for them to perceive the larger picture and become aware of the interactions.

Search is currently studied intensely under the name of *information retrieval* (IR), by academics and industrial researchers alike. I have consciously avoided this label in the title of this thesis and its various chapters in order to indicate a focus on the behavioural phenomenon of searching rather than the algorithmic problems involved.

Information retrieval research being considered a non-humanistic field, there is a feeling among some less philosophically minded researchers that all research should be based on empirical studies and that preferably no conclusions should be made without reference to such studies. However, I believe that the study of search is in fact largely a humanistic science by nature, and that consequently these – often fruitful – methods need to be complemented by a more dialectic approach. Some things simply cannot be measured, at least not at present, so we must allow for some degree of introspection and philosophical speculation, as long as it is founded on solid argumentation and complemented by convincing exemplification. Such introspection combined with analysis and synthesis of observed data is in fact considered an established and successful method of constructing models in some theoretical branches of the field and is hence adopted here (Ingwersen & Järvelin, 2005a). The Stage-based Intent Model developed in Part III with the aim of explaining search behaviour and query reference is one example where reference to existing studies and the data provided by the analysed search logs have been supplemented by pure introspection in order to arrive at a model that at least is able to describe and to some extent even explain the observable behaviour. There is no direct proof to be had that such a model is correct, but if some data are found to contradict the model, one will merely have to adjust it, rather than starting from scratch every time. Not least, having a framework makes it possible to discuss its features in the first place, which in itself is a huge advantage. The Regulated Flux model of relevance suggested in Part II is based on relevance evaluation data published in previous research papers by other authors but interpreted in a new way that is suggested to be far more telling and suitable for use in the construction of an explanatory model. In this case the very empirical method that was meant to ensure reliability has in fact contributed to the misconception of what the real problem is because of failing to reach the level of abstraction at which the phenomenon of relevance operates. Researchers record user reactions, but the reactions do not represent the input to the relevance mechanism; rather, they cause certain psychological states to arise in the users which secondarily affect relevance assessment. Thus, what cannot be measured can be identified experimentally, but it is important not to believe that what such experiments show is the reality; that will often have to be reconstructed indirectly through abstract analysis. The two models mentioned here fit beautifully together via an intermediating cognitive model of facets that applies to topics and resources as well as situations perceived and experienced by the user – the Faceted Stimulus model introduced in Part I.

The material for the empirical analyses of search behaviour and language use drawn upon in this thesis comes from enterprise search in a wide sense, i.e. search performed on smaller document sets such as a single web site or a public collection of resources. I have had the immense privilege of working for the Danish language technology firm Ankiro alongside my studies, and this has opened my eyes to the inevitable gorge between academic and business research. One has to try a bit of both in order to understand the true nature of the difference. Much of what academia comes up with – including the present thesis, which is, after all, highly academic in nature – is not immediately applicable to the needs of even a research-minded technology provider such as Ankiro. Research results always need to be tempered with the pragmatics of what is possible in actual working systems at the present time, and this requires sacrifices on both sides. The inescapable fact, however – and the reason that I have been asked to carry out this research – is that visionary and horizon-widening studies of the most academic kind imaginable are a prerequisite for any real progress that will make the impossible achievable in future. This thesis is a cross between an industrial and an academic PhD in that it focuses on treating some highly theoretical issues while at the same time suggesting solutions to a range of minor problems met with along the way which might puzzle the system designer. With an industrial solution in mind, no stone should be left unturned, even though only a fraction of the plentiful questions that arise can be answered at present. Hopefully, this is not too much of a stumbling block to the academically minded reader. The IR field currently receives a lot of interest from many sides – academic as well as business-oriented – and consequently spans a huge number of theories, many of which are mutually incompatible. It is my hope that this contribution will help the field advance on a somewhat safer ground in future.

Acknowledgements

Before setting out on this daunting yet fascinating journey I would like to thank Ankiro for giving me the unique opportunity of delving into the engaging matters of linguistics in search, and for trusting me so wholeheartedly as to give me free reins in my research. Creativity is a, sadly often neglected, key part of research and should be nourished accordingly. Next, I want to thank my supervisors at Copenhagen Business School for bearing over with my eccentricities and somewhat atypical approach to research.

Working on this project would not have been half so enjoyable without the continuous support and encouragement provided by a range of people around me. I would consequently like to thank my colleagues at Ankiro who keep encouraging me in my work and always make me feel so much at home in their company. Research always improves with the amount of time spent in real life, and working on abstract theories one day and applying them in real-life search engine improvement the next has been a rare treat.

I take this opportunity also to thank TDC, the National Danish Labour Market Authority and Novozymes for letting me use their search data for the empirical analyses. Furthermore, I want to thank my students and colleagues with whom I have had fruitful and interesting discussions. I would also like to mention the lexicographers Henning Bergenholtz and Patrick Leroyer of Aarhus University and Theo Bothma of University of Pretoria who opened my eyes to lexicography and shared their considerable knowledge with me, and to thank Roman Yangarber and his skilful team in the Department of Computer Science at Helsinki University for including me so warmly in their team and letting me partake in their work. Finally, warm thanks naturally go to my wife for her endless patience and support, and for following me around the globe in my quest for understanding.

Esben Alfort

Central concepts

This overview includes short explanations of some of the central concepts introduced in this thesis. The list is for reference only; the reader is referred to the text for more information.

- ***The facets of a resource***

The intentional resource (I): The resource as an object intended for a particular (set of) recipient(s). This is an abstract facet of the resource comparable to a *gift* considered independently of the thing given.

The substantial resource (II): The resource as a physical (or digital) object to be accessed by someone. In a textual document, this facet includes raw strings with no linguistic sense ascribed.

The formal resource (III): The resource as an object with an intrinsic value and form. In a textual document, this facet includes linguistic terms.

The representational resource (IV): The resource as an object of a certain type with certain uses. In a textual document, this facet includes the semantic content.

The interpersonal resource (V): The resource as an object conveying a message or fulfilling a purpose. This facet is similar to a *resource* that one may draw on, irrespective of what it consists of.

The effective resource (VI): The resource as an object with a cognitive or psychological effect on its receiver(s). This facet includes the aesthetic qualities of a document but also its informative qualities.

● ***Contextual spheres of relevance***

Focal relevance: Something is selected by the individual as being of current interest. In search, this is the topic(s) represented in a query.

Local relevance: Something is perceived as being problematic and is distinguished from the horizon as a theme on which to concentrate. This includes considerations of utility in a given situation.

General relevance: Something is judged as being relevant on the background of the horizon, the background knowledge possessed by the individual and the past experiences. This includes what is merely of general interest to a person with no reference to the current situation.

● ***Dimensions of relevance***

Intentionality (I): Was the resource intended to be read by the user or someone like the user?

Practicality (II): Can the resource be accessed by the user in a readable format and read in a language known by the user?

Informativity (III): Is the information content able to change the user's understanding and provide cognitive gains?

Topicality (IV): Is the content on a topic that is currently considered relevant by the user according to the Hierarchy of Relevant Contexts (i.e. the contextual spheres of relevance)?

Utility (V): Is the resource helpful in performing a task (including the simple collection of information)?

Desirability (VI): Does the user consider the resource satisfactory and appealing?

● ***Referential targets***

Motivation (C1): The affective drive towards a possible personal gain that triggers the user to approach a search engine (or other linguistically activated tool) in the first place.

Source (S1): The problematic situation or sense of lack that makes the user realize that a search needs to be made.

Trigger (N1): The actual need that triggers the user to search for something particular.

Subject (T1): The topic that the user chooses to search for in order to reach desired resources.

Object (T2): The actual ideal resources that the user has in mind.

Aim (N2): The ideal problem-solving outcome of the interaction with the (extended) system that the user envisages.

Goal (S2): The ideal, problem-free state that the user desires to reach through interacting with the (extended) system.

Objective (C2): The ultimate purpose of searching characterized by personal gain or new abilities due to the (imagined) successful search.

Part I

Studying search

Chapters 1-3

1 Introduction

> *The concern is that in the present information environment rapid, unrelenting transmission of information has diminished the prospect of synthesis and the satisfying whole experience that comes from deep understanding. The challenge is that of designing information systems and services which enable people to move from uncertainty to understanding in an information world where one is threatened with the prospect of being in a perpetual state of uncertainty.*
>
> (Kuhlthau, 1993)

By now sufficiently many have pointed out that the tremendous amounts of information that surround us are such that we cannot possibly cope with it all for it to have become a truism. Thus, the following quote will not in itself surprise the reader in the slightest.

> Everyday living is too fast, too busy, too complicated. More than at any time in history, it's important to have good information on just about every aspect of life. And, there is more information available than ever before. Too much in fact. There is simply no time for people to gather and absorb the information they need.

It could have been uttered by any modern observer, but interestingly the words are in fact those of Britton Hadden shortly before he co-founded Time Magazine in 1929 (Delphi Group, 2004). The problem is thus far from new, and has only been growing in magnitude, but even so it remains unsolved; we are continually bombarded with news and instructions, and information is forced upon us as soon as we enter a cyberspace loaded with resources just waiting for us to retrieve them, if only we know what we are looking for. However, finding what we do want or need in this immense informational swamp often proves a taxing and overwhelming task. The vast majority of the information that we encounter is irrelevant in the sense of satisfying our immediate needs; so much of it needs to be filtered away or ignored that it is easy to bypass the vital or even to be sucked down and give up entirely, a condition which in Scandinavia has come to be known by the rather dramatic term *information death* (Bergenholtz & Gouws, 2010).

We regularly communicate our problems, needs and desires to machines such as search engines, and we expect them to "do their best" at returning useful answers or at least relevant resources. Helping users optimally and in as short a time span as possible is in fact not the ultimate *raison-d'être* of all such systems, due to the powerful economic forces involved, but even so most **information retrieval** (**IR**) researchers, at least within the fuzzy boundaries of academia, choose to adopt this naïve but reasonable assumption in an attempt to identify better ways of retrieving exactly what the users of such systems need.

> The objective is to study and understand IR processes in order to design, build and test retrieval systems that may facilitate the effective communication of desired information between human generator [i.e., author] and human user. (Ingwersen, 1992, p. 49)

Some providers of information exist for this purpose alone, and many years of working with such customers have taught me that they really do want to help their end-users by all means possible. More and more sophisticated systems have had to be developed for their benefit in order to cater for ever more specific types of request, and this demands an increasing degree of awareness and understanding as regards people's needs and ways of expressing them. However, the communication of needs is typically conveyed through the medium of some form of language, and language is notoriously difficult to deal with computationally. For this reason, continued research into the matter is crucial. Thus, while mainstream web search researchers have jumped off and boarded their own train down the road of statistical methods for the benefit of the masses, it is time that enterprise search researchers constructed their own vehicle and took up the task of finding a way to better understand queries, where the mainstream research field left it years ago. It is not the case that no such research is taking place, but progress is impeded by the fact that there is no unified field and no overall theoretical framework to refer to. Partly, this is due to the secrecy inherent in commercial research, but it is also very much a result of lacking foundations on which to build. Such a unifying framework is what I hope to be able to provide, or at least contribute to, through this work.

If we are to help users, we need on the one hand to understand their problems, needs and wants, and on the other to draw up a model of the relationship between these and the way users express themselves through queries. Neither of these subjects is properly understood today. Some researchers, especially within the library tradition, focus on incorporating a fuller description of the user's perceived problem, while others are more interested in identifying users' needs and wants, generally through interpreting queries more fully. However, queries may in fact describe either problem states or the goals and desires of users, as we shall be seeing below. What makes users choose one of these strategies, and how the expression of the query relates to a particular need is simply not known at present.

Solving this problem requires modelling the choices that people have when formulating a query and attempting to relate this to actual surface expressions. This will help us understand how to interpret queries in order to automatically construct a user profile containing information on the user's problem state and/or desired resources. These are the contributions that we shall be attempting to make at present. The next step will be firstly to extract and disambiguate the information from queries at run-time, secondly to develop algorithms that exploit the information as part of the retrieval mechanism, and thirdly to index documents accordingly with the appropriate metadata that will allow a search system to retrieve the exact resources that match the needs and wants *conveyed* by a certain query (obviously, what is not conveyed will usually have to remain vague). These are topics that future research will have to deal with once an understanding of the problems we are facing has been achieved.

There was a time when search engines just looked up the terms in your query and returned links or references to any documents that contained the words in question, but by now there is so much information out there that most queries would deliver unmanageable sets of results if such simple algorithms were used; some sorting must occur before presenting the results to the user. Also, while it used to be possible to refind a certain webpage months later by entering the same query again, nowadays search engines are so dynamic and flexible, monitoring your every move and listening to tendencies and preferences, that every search result is slightly different from the one before, even if the query is kept unchanged. Thus today, what you ask for is not what you get. The more search engine designers attempt to capture the subjective needs and interests of the individual, the more they do in fact rely on general tendencies and the averaged preferences of the multitudes. Now, the query has largely been demoted to a mere guideline, the input to a statistical analysis of what other users, who are presumed to be similar, seem to prefer.

Many people are starting to find all this a little unsettling. It is, however, likely to be a necessary development when dealing with such an extensive and varied collection of source documents as the World Wide Web, unless a radically new method is developed. Maybe such entirely new methods will be required in the near future, as the pace of growth accelerates (Blair, 2002a). This calls for more and more restrictive algorithms, as the amount of unwanted noise escalates. In smaller collections such as individual homepages or document collections that allow users to search for information within their limits, it might just be possible to keep listening to what the user is actually trying to communicate and attempt to understand his or her needs directly instead of falling back on sterile statistics. In fact, in a collection of more reasonable size such methods would be likely to outperform even the most sophisticated statistical techniques, since the effectiveness of the latter is limited by the fact that they only work on large datasets. The present research is thus geared towards search in smaller environments in the first hand. However, even the larger collections may benefit from this kind of research because, as suggested by Blair, it may be a good idea in future to deal with the enormous size of web collections by performing *partitioned search*, where the collection is first reduced to a manageable size by filtering away clearly irrelevant material before activating a second algorithm that focuses the search on what is relevant. Thus, we may find ourselves using more and more of these linguistically and cognitively informed algorithms even in web search in future.

1.1 Research goals

The ultimate objective of this research is to explore whether search tools for collections of limited size can be made to understand what a user is trying to communicate in order to improve search results and help users through the process of satisfying their needs. The central question that presents itself for investigation in this connection is what determines how a user chooses to phrase a query, or in other words, what factors affect the choice of words, morphological form and syntactic organization. Only when we understand this can we hope to be able to "decode" users' intentions from the queries that they issue.

What determines how a user chooses to phrase a query?

This question is highly interesting not only from the point of view of IR research, but also for theoretical linguistics, cognition and relevance studies, and it is hoped that the present research will incidentally in its own way contribute to a better understanding of the linguistic and cognitive topics covered. After all, we do not really know what makes people focus on certain aspects of reality and include them in a linguistic portrayal while simply ignoring huge amounts of information also available to them. Only some aspects are considered *relevant* in a given situation, but exactly what the notion of *relevance* involves is a contentious issue.

In order that we may get closer to an answer to our central question, this thesis explores language use, need development, search behaviour and relevance assessment. We shall examine a range of ways in which queries may vary, such as by their morphology, syntax, word choice, level of generality, referential choices, and linguistic function; the kind of variation that makes dealing with language so difficult. While the semantics of the words contained in a query may describe the need or the problem or the desired end state, etc., it is the choices that users make when deciding (more or less consciously) how to express themselves that really signal their intentions. Thus, in this thesis we shall focus on the question of what users choose to refer to in the first place, given a choice.

What do queries refer to?

Imagine that you were looking for a new job. Would you describe your qualifications, or the kind of job that you were interested in? Similarly, if your USB drive was broken, would you be more likely to search for help using the query *usb broken* or *usb repair*? Both solutions occur quite frequently. In the USB case, the interpretation does not present insurmountable difficulties, as the two queries require similar responses, but how could a job search portal possibly determine whether *phd copenhagen* means that the user is a PhD from Copenhagen looking for a job, or someone with a master's degree living somewhere else but looking for a position as a PhD student in Copenhagen?

Clearly, questions such as these require some looking into. The first thing to do is to develop a model of what users may possibly refer to with their queries. Only when this is completed will a discussion of what particular users actually choose make sense. The main goal is thus to provide a framework for understanding and analysing queries. This is achieved through five partial aims, all of which are closely related.

1) Developing a model that is able to explain the immense variety of interpretation. What is the relationship between a document and its contents, or between a situation and the task that it centres around? How do affective issues and relationships between people enter the equation?

Part I introduces the Faceted Stimulus model which acknowledges the fact that resources have many facets that may be judged separately by the user. A document is simultaneously a physical (or digital) entity, a semantic content with uses in a range of situations and a piece of communication from an author to a reader, and it is important to keep these facets apart while at the same time remembering to take them all into account when considering system design.

2) Developing a model that is able to explain users' relevance assessment of resources, given certain needs in context.

The question of the nature and characteristics of relevance has long been considered one of the great mysteries of IR research. Everyone is able to use the concept in practice, but explaining what it is and how it works proves extremely challenging. Part II is an attempt at improving our understanding of this engaging phenomenon as a first step towards understanding the (often subconscious) reasoning behind search behaviour. The Regulated Flux model developed in Part II helps understand and potentially even algorithmize the complex processes involved in relevance assessment and the way they intertwine and interact.

3) Developing a model that helps understand why users express their needs in certain ways.

Part III contains an introduction to the Stage-based Intent model which enumerates and relates the various choices that users have when deciding how to express themselves in a query. Understanding user expression involves a combination of cognitive and linguistic knowledge and awareness of relevance and applicability issues.

Building these three models is the core aim of the thesis. However, a fourth aim presents itself in this connection.

4) Providing a theoretical discussion of the linguistic issues involved when users express their needs, i.e. identifying types of morphological, syntactic, lexical and other variation in query formulation, and attempting explanations of this behaviour.

The empirical data necessary for such analyses are drawn from search logs, but log data are notoriously difficult to make sense of because queries are ambiguous and short and leave out a lot of contextual information. No satisfactory method for analysis of search logs has been forthcoming, but the development of the above-mentioned Stage-based Intent model of user expression promises a possibility of actually developing a method based on the suggested concepts, and this shall be our fifth aim.

5) Developing a method for analysing query logs that derives its data from the linguistic expression of queries alone and thus allows for unbiased classification.

This research is not purely academically motivated; it arises from practical needs in the enterprise search field, and so this other task begs to be fulfilled even though it is not strictly part of the overall goal as defined above. Even from an academic point of view, however, it is only natural to provide a method based on the principles developed here as an example of how to apply the theory in practice. The need of finding a method of interpreting query logs is one that is ever present for any professional IR systems developer, and certainly one that I have felt in my own work. The method suggested here has by now been tested and used commercially, but it must be said that it is not likely to be quite as easily applied to just any domain; the data on which it is applied here is probably more easily applicable to it than most, primarily because of its high level of homogeneity, but also because of the nature of the domain itself.

To the above goals must be added the secondary – but hardly any less difficult or controversial – matters of exploring and modelling the ways in which needs develop and determining what contextual parameters are necessary parts of the model. Furthermore, in order to facilitate scientific discussions of query variation it has been found invaluable to have a formal representation of queries that reflects the understanding provided by the suggested model. With such a representation, it will be possible to notate the two interpretations of *phd copenhagen* in different ways as referring to what we shall be calling the *Source* and the *Goal*, respectively[2].

1.1	**phd copenhagen** {S1/S2: ($phd_{n.sg}$) > ($copenhagen_{n.prop}$)}

Also, the fact that the terms *phd* and *copenhagen* refer to different entities becomes clear with this notation.

Hardly any of the enumerated steps could be left out without losing the power of the overarching framework and the understanding that it provides. Many previous studies have focused on smaller subsections of the problem and have provided important data, but it is essential at this point to zoom out and embrace the entire phenomenon as far as possible in order not to aim in the wrong direction. It

[2] Sources (S1) and Goals (S2) are explained properly in Part III. In this particular example the difference involves describing either the situation that existed when the need arose (and may sometimes have caused it) or the one that will satisfy it.

is of course inevitable in such a project that some important details may have been given too little attention. The reader is therefore encouraged to consult the literature on the various topics.

1.1.1 Assumptions

Queries are commonly treated in IR studies as complete, immutable strings that can be read from search logs as objective data[3]. This enables researchers to study them as the finished input received by the system. In this view, some queries are classified as representing predictable user goals and others as being unpredictable, and the latter are treated as if they were simply instances of imperfect data, i.e. had the query not been ambiguous, assigning a certain intention to it would have been possible.

> A user's goal for a query is inherently subjective. Thus, the first question is whether it is ever possible to associate a query with a particular goal simply by looking at the query without any user feedback. For example, our user study shows that most users associate the query *bestbuy* with the official BestBuy Web site and consider the query navigational, while the user opinion on the query *Alan Kay* is evenly split. Some people want to visit the homepage of Alan Kay, while others want to read multiple pages related to Alan Kay in order to learn about his career and research given his recent reception of Turing Award. When the user opinion is evenly split, it will be clearly difficult for a search engine to reliably predict the goal of a user without collecting any further information from that user. Given the above sample queries, it will be highly interesting to study how many queries will have a predictable goal, and how many queries will be "unpredictable" in their goals and require further information from the user for reliable prediction. (Lee, Liu, & Cho, 2005, p. 391)

This distinction between predictable and unpredictable goals completely bypasses the human cognitive aspects of query production; a query does not necessarily represent a particular intention or goal, it may simply be inherently vague.

Ex.	Query	Translation	Log (*n*)[4]
1.2	**medarbejdere**	'employees'	TDC (85)

[3] I.e., objective data of subjective phenomena.
[4] Examples from the analysed log files from TDC and Workindenmark are marked *TDC* and *WID*, respectively. Made-up queries included for illustration are given an asterisk. The figure indicates the number of times this query occurs in the log in question. See chapter 3 for more information on the search logs used.

The users who issued this query may not all have had either an intention of locating a list of employees or of retrieving information on employee conditions; they may have wanted both simultaneously, or may not have had any opinion or preference in this respect. As pointed out by Winograd and Flores (1986), intentions do not have to match what is achievable in the real world because the human mind largely works in an irrational way.

As is clear from the above quote, researchers may well be aware that queries are subjective, but many still carry out their investigations as if they were not. While this approach is suitable for the development of new search algorithms and studies of search engine effectiveness, it ignores the psychological search process and hence misses the opportunity to come to understand what makes people phrase their queries the way they do. In this thesis, queries shall be interpreted as dynamic products of cognitive processes which may reveal clues to the user's needs that are commonly overlooked at present. The approach adopted here is thus born out of the cognitive approach to IR (Ingwersen, 1992; Ingwersen & Järvelin, 2005a). The approach implies that information is considered to be constructed by the reader rather than being objectively present in a document, that tasks in the external world are seen as central in the need development process, that contextual information is essential for the interpretation of queries and requests, that interaction is at the centre of search and is what the researcher must pay attention to more than anything, and that some kind of intermediary component is required[5] to mediate between the user and the system in order to translate human requests into something understandable by the system, whether simple or complex.

It is regularly pointed out that queries are extremely short and therefore reveal very little about user intentions through their linguistic make-up anyway. To a certain extent, this is true. Most queries certainly are very short and often ambiguous (cf. section 3.2 on query length), but much that is ambiguous in isolation becomes clear in context. When speaking, people rely on their interlocutors to be able to infer much more from what they say than is actually expressed in words, and it is a basic assumption here that this behaviour is so deeply ingrained that we cannot refrain from expressing ourselves in similarly context-dependent ways even when the interlocutor is a machine and hence cannot be expected to be able to perform the necessary inferences. The hope is that we may indeed enable search systems to access some of this information, or at least exploit what *is* said in an optimal way by attending to variations in query expression and the differences in meaning that go with such variation.

[5] Or at least is highly desirable.

That communication involves much which is never said is hardly a new insight; it was first demonstrated by the British philosopher Herbert Paul Grice (1975), and is by now a widely accepted truth. However, the fact that it applies to queries as to other instances of language seems unfortunately often to be forgotten (Ingwersen, 1996; Blair, 2003), or perhaps it is conveniently ignored due to the inherent difficulties of dealing with context in an appropriate way.

> Context is one of the most important concepts in information seeking and retrieval research. However, the challenges of studying context are great; thus, it is more common for researchers to use context as a post hoc explanatory factor, rather than as a concept that drives inquiry. (Kelly D. , 2006, p. 1729)

This is not to say, of course, that context as such is ignored – quite on the contrary – but the fact that it influences language use in queries is not part of most contextualized models. Rather, context is taken to primarily affect user needs and resource text sense. Context is dealt with here in section 2.4.

1.1.2 Methodological considerations

In most empirical studies of IR systems, researchers examine the set of search results retrieved in response to a given query in order to assess whether the resources in question were relevant to the user in the particular context that prompted the search (Ingwersen, 1996). This manner of looking at the search process is not suitable for a study of expressive variation in queries such as we shall be conducting here, due to the analytical focus on variability of search results rather than of query form. Here, we shall not be interested in what the search results would be given a certain query, but rather what query a user would choose given a certain desired result, which is the opposite point of view.

There are three main parameters in the equation that we shall want to examine.

1) The documents *D* sought
2) The query *Q* used to find them
3) The context *C* in which the search is performed

Even though the user mostly cannot know what documents are available, *D* is here taken to be an essentially static (though hypothetical) collection of resources answering the user's needs, i.e. *D* is not the actual collection searched or even retrieved, but rather the set of potential documents which would ideally satisfy the user's needs, whatever they might be.

There are three ways of studying the relations between *Q*, *C* and *D*.

- **Relevance and IR effectiveness studies**: Most information retrieval researchers investigate what documents *D* are relevant to *Q* in the context *C*, because this mirrors the task of a search engine, which is to identify *D* based on *Q* and, as far as possible, *C*. In this model, anything subjective is assigned to context, while the query is considered static and invariable. Even though this is never stated explicitly, the model causes users to be considered more or less as slaves of their context; they are effectively not given a choice as to how to phrase their queries, because any variation in choice is given contextual explanations. It is true that the user's mental considerations are part of the context, but even so, this model studies the way users evaluate the relevance of documents and not the way they search for them. The method does allow different queries to be compared, and their varying effects can be studied across various contexts, but subjectivity inevitably becomes a peripheral modification of the objective relevance derived from a particular context. We shall therefore need to take a different perspective. This is not to say, of course, that it is not the right way of studying relevance and IR effectiveness; only, that it is not what interests us here (except of course when dealing specifically with relevance in Part II).
- **Context and need studies**: Alternatively, it is possible to study in what context *C* a user submitting the query *Q* would be interested in the documents *D*. Such a model would study user needs, but not the way they are satisfied. Consequently, this is not the setup that we shall be investigating either (except briefly in our discussions of need development and context below).
- **Query expression studies**: In order to acquire a better understanding of users and their choices, we shall in the main choose a third viewpoint and examine what query *Q* a user in context *C* might utilize with the aim of locating the relevant resources *D*. In this way, the user's choice moves centre-stage, and the study focuses on the user's (often subconscious) reasons for phrasing a query in a certain way; this is a study of user behaviour in search, rather than relevance judgment or need origins.

Apart from the question of what parameter to focus on, the choice of setup also has repercussions for the values that the variables may take on, because any static parameter must in the experimental setting have a finite and hence predefined set of values. Otherwise, the experimental setup itself cannot be said to be fully understood. Hence, a preconceived theory must be applied to such parameters. In contrast to this, the third variable can be evaluated on an analogue scale. The result is a richer description of the dependent focus variable.

The two first setups described above keep the query static, and accordingly only a predefined set of search strings can be considered. Typically, a set of test queries are selected either from a log file or from a test collection, and then a search is performed. The retrieved documents are evaluated as to relevance and the results are claimed to be in principle generalizable to other queries[6].

If the focus is on the document collection as in the first setup, *C* and *Q* must be static, and *Q* is selected from the set of queries that a user in context *C* might use with the hope of retrieving relevant documents. Any document result is possible given a previously chosen combination of *C* and *Q*. *D* is thus an analogue value representing whatever documents are relevant to the user in context *C*[7].

If the focus is transferred to the context as in the second setup, *Q* and *D* must be static. *Q* is selected from the set of queries that a user might employ in order to find *D* (according to some set of hypotheses), and *D* is selected from the set of document collections to which *Q* might possibly refer. *C* is a complex, analogue value representing any context that answers the studied situation.

When the focus is on the query *Q*, as is the case in this thesis, *C* and *D* must be kept static. *C* is selected from the set of context types in which a user might want to find *D*, and *D* is selected from the set of document collections that a user in the context *C* might be interested in finding. Thus, the two are interdependent constructions of the experiment. *Q* is an analogue value, i.e., any expression is possible. This is a new way of doing IR research, and one which I believe will prove a fruitful complement to traditional approaches. It implies that the context of a query must be known and classifiable to a certain extent, as must the user's information need. Given that our data consist of search logs, the knowledge of context is naturally limited to what we know about the sites on which the searches were performed, but this can in fact tell us a lot about the situations of use. The range of possible needs is inferred on the basis of a theoretical framework and one at a time is chosen before compared to the actual expressions used to express them. This allows the analyst to control the unavoidable ambiguity.

[6] These methods have been amply criticized through the years. One simple, valid point of critique is that it is impossible to scale results from experiments made on small collections to larger ones, because they do not behave similarly, for linguistic reasons (Blair, 2002a).
[7] Note that *D* is not even constrained by the resource collection searched, as *D* represents ideal, mentally represented documents and not actual retrieved ones.

1.1.3 Defining central notions

The fact that in this kind of research the two controlled variables are heavily interdependent might initially seem problematic. After all, how can it be reasonable to define *C* based on *D and vice versa*? The answer to this question has to do with the fact that a need is a relation between a context (*C*) and a document set (*D*). It may be defined for our purposes in the following way.

> A <u>need</u> is a relation $n(D,C)$ between a context C and the resources D that are relevant to the user $u(C)$[8] in the context C. The relation is such that D comprises exactly those resources needed by the user in context C.

Thus, what is constructed for the purposes of our analyses is actually a **need** $n(D,C)$, i.e. it is a single relation rather than two independent variables. This would seem to make the endeavour reasonable. Needs are typically conceived as relations between a user and some information, and this is reasonable from the point of view of behavioural studies, but from the system's point of view the individual searcher is merely part of the context that creates the need as well as the factor that defines what resources are needed; to the system, there is just a need-generating context and some needed resources, but the user still pervades the definition by shaping both *C* and *D*. The fact that there is someone actually *experiencing* the need is thus not part of the relation itself, in spite of this individual's obvious importance. A similar explanation can be made for the typical setup employed in IR research, where *C* and *Q* are constructed. In this case, the pair may again be considered a single relation $r(C,Q)$ representing a **request**.

> A <u>request</u> is a relation $r(C,Q)$ between a context C and a query Q issued by the user in the context C. The relation is such that C comprises exactly those aspects of context which are relevant for the formulation of Q.

Once again, we shall not choose to refer to an individual in the definition of a request that we shall adopt. Rather, a query is issued in a context defined by the user's view of the world, and this is what matters to the system. The fact that someone is *making* the request is an interesting point but is not part of the relation itself. Finally, in context and need studies, the constructed relation $s(Q,D)$ may reasonably be called a **search**.

---

[8] As discussed below, a user is a function of the context because context determines what the user needs and this is the part of the user that an IR system cares about. Thus, *user* as understood here differs from *searching individual*.

A <u>search</u> is a relation $s(Q,D)$ between a query Q and the resources D that are relevant to the user issuing Q in the context C. The relation is such that Q reflects the need for the document set D.

A search as defined here is thus not performed by a user, not even in the sense of a context shaped by the user's consciousness as was the case in the other two relations. Rather, it occurs between a query and a set of desired resources. This, then, is where the user enters this last relation; the resources *D* are defined by the user's needs and desires. Summing up, the three types of IR research compare the variables and relations in fig. 1.1.

	Static relation (Bound variable)	**Free variable**
Relevance and IR eff. Studies	The request $r(C,Q)$	The resources D
Context and need studies	The search $s(Q,D)$	The context C
Query expression studies	The need $n(D,C)$	The query Q

Fig. 1.1. IR research setups

The order in which the variables are mentioned here is not coincidental. For instance, it is not the desired or needed document set *D* that reflects the query but the query that should reflect *D* if the search is to be successful. One might even say that it reflects *D* by definition, though the degree of fit varies depending on the user and the situation. This is a really important choice in assumptions, because it represents the difference from an algorithmic study of how actual document sets vary with changing queries to how queries are formed depending on the desired resources. In other words, by considering *Q* to be derived from *D* rather than *vice versa*, what could have been a study of systems and algorithms is turned into a behavioural linguistic investigation.

The same choice holds of the other two parameters. A query (or any other expression for that matter) creates its own context, because it is only when there is an expression that there is something for the context to be a context of. Consequently, the expression is not considered a product of the context, but the other way round. This means that instead of studying how queries arise in certain contextual settings, we study how psychological contexts vary with the queries that they are the contexts of. Looking at what parts of context are made relevant by different queries can tell us much about the intention behind the queries. Once again, this makes it a behavioural study rather than a transformational one.

Finally, context dictates what the desired resources are. Rather than trying to describe a need in contextual terms, we shall therefore be studying how context affects the human mind and generates a need. A problematic situation does not only create a need, it also drives user behaviour, which may shape the need in various ways so that the user ends up looking for something entirely different. In other words, what is important if you want to understand user behaviour and need development is not just the initial situation but the way it affects the human mind and behaviour which in turn leads the user to settle on a need.

If *Q* defines *C*, *C* defines *D* and *D* defines *Q*, there is equally a relationship in the opposite direction, though not one of definition (fig. 1.2). It is clear that *previous* context affects the formulation of the query, and that the desired resources form part of the *ensuing* context by representing the wished-for outcome. Similarly, *Q* may indeed influence the actual result, and this in turn has the potential of modifying the user's needs (by affecting the user's understanding of the problem), and this is equivalent to a subsequent change in *D*. The new ideal document set again modifies the context and so on and so forth. The processual nature of search becomes very clear with this circular model.

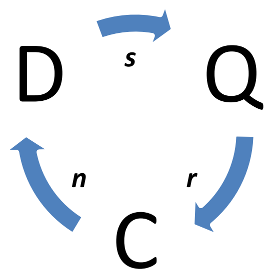


Fig. 1.2. The relations between context *C*, query *Q*, and ideal resources *D*.

The traditional view of the search process is that searches are performed by an individual called the *user* in a *context*. The context is defined from the user and in turn affects search behaviour and relevance assessment. However, we shall contend that the individual user as such does not have any importance for the search process. What matters in retrieving relevant resources is the user's need, which is defined by the user context as perceived by the individual. In fact, it makes sense to consider this perspectivized context as in some sense *being* the user, as by definition it constitutes exactly those parts of the user's mind that are relevant to the search. In light of this understanding, **users** (in our IR terms) are not only functions of their context, but of their needs, which in turn depend on context and ideal document sets. In fact, a user may properly be defined as the combination of an associated request *r*, search *s* and need *n* (fig. 1.3).

A <u>user</u> is defined here as the combination of an associated request $r(C,Q)$, search $s(Q,D)$, and need $n(D,C)$.

This is not to say, of course, that a definition based on the individual is not appropriate in other types of studies, but from the system's point of view there is no individual, only a user represented by the combination of an associated request, search and need as defined above. Indeed, the individual performing the search may not be the user who has the need and for whom the context is valid; it is entirely possible to search on the behalf of others, or as oneself at a different point in time from the present.

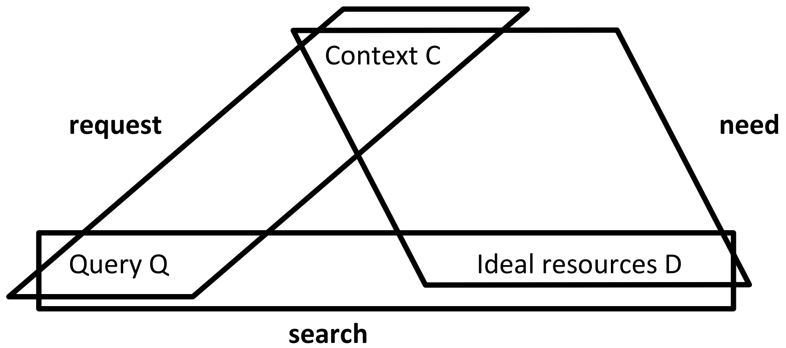


Fig. 1.3. The three constituents of a user as technically defined here.

Our definition implies that each time one of the parameters involved is altered it is by definition a new user. Context changes significantly between searches, as responses from a previous search act, and even the user's own reflections, immediately become part of it. This gives emphasis to the fact that users are a product of their context and change with it, an important insight of previous IR research (Morris, 1994). Of course, it is often useful to be able to state that the same **individual** makes several requests, performs a series of searches, and has various needs, but it is probably advisable to refrain from ascribing this to the same *user*, at least in the kind of studies performed here. Many authors consider the user to be the same across a **session** – a series of connected searches performed with the same overall goal in mind (Xiang, et al., 2010). Since the Target Analysis method developed here is so far a manual technique, the large search logs have had to be reduced to a small percentage of their original size when needed for such analysis, following a procedure described below. This reduction makes session analysis impossible because usually only a single query will have been included in the final dataset from any given session. The adopted definition of *user* does not preclude identification of consecutive queries issued by the same user, or even during all the searches in an individual's experience. Such facts are simply part of the user background; insofar as earlier queries are still remembered, they are still part of the user's experience and mental state. Considering the user to be ultimately a function of context has the advantage of allowing the changeability of the human mind to be incorporated into a model of search in a natural way. Similarly, a certain string entered in a search box repeatedly is a new **query** every time it is submitted; the same string as utilized many times by different users represents different queries even though the intention may potentially be the same. The fact that some strings are more popular than others may be captured as part of the context of the search.

Given the definition of *user* adopted here, the reader will appreciate why this thesis must contain a study of both relevance (i.e. of the ideal document set) and context as well as of queries. Note, however, that with the given definition a query is not a function of context and desired document set; a need is. A query therefore cannot be predicted from context and desired document set alone, as there are no clear translation algorithms between need, search and request. What we are doing in this thesis to solve this conundrum is studying the relationship between queries and needs, thus effectively making a stab at providing one such translation.

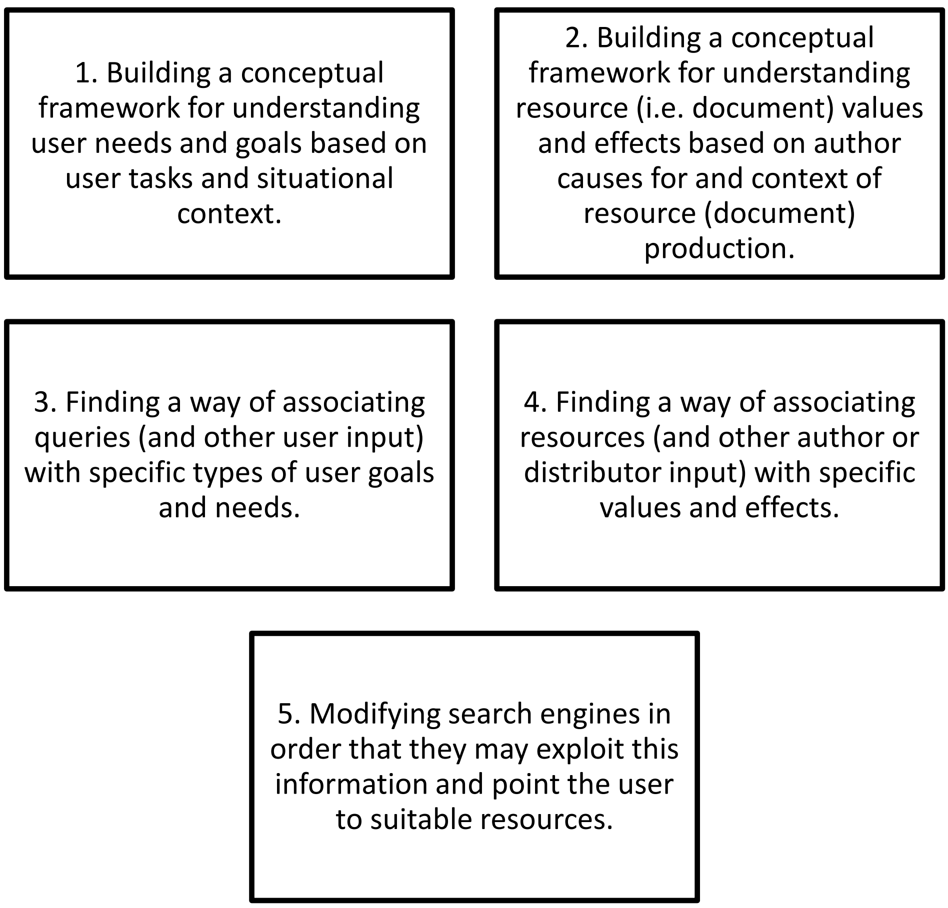


Fig. 1.4. Schematic overview of the work process needed for improving information retrieval. This list is an expanded version of the one suggested by Rose and Levinson (2004), who do not consider the resource side and consequently have only three tasks on theirs.

1.1.4 Prerequisites for success

In order to reach the goal of improving the capabilities of search systems to match user intentions and needs to the informative value of resources and author intentions, at least five steps need to be gone through (fig. 1.4). The focus of this thesis will be mainly on step 1-3. Step 1 involves looking closely at tasks, motivations, needs, goals and aims. This will be one of the major themes. Step 2 requires research into the causes for text production and the semantics of documents, which is only touched upon briefly here. However, the notion of relevance combines these two subjects, and this is a notion that is discussed extensively in Part II. Step 3 involves studying queries and their relationship to user intentions. This is ultimately the main theme of the thesis. Step 4, on the other hand, requires semantic and pragmatic analysis of documents and/or meta-tagging. Identifying a user's intentions is not very helpful if the informative value of the resource documents is not also known. For instance, even if a system succeeds in determining that a certain user wants to buy a car rather than merely gathering information on cars of a certain make, he or she can only be helped if the system also knows what sites actually offer to sell cars rather than merely providing information on them. Apart from semantic information, documents accordingly ideally need to be indexed and tagged with the document author's communicative intentions, or at the very least the type of site that they belong to, in order that people may find the kind of information or service that they are looking for. Like many of its predecessors, this thesis focuses on identifying the intention of the user and basically assumes that a solution to the indexing problem will be found while it is left to a separate research field to take care of the optimization of indexing methods. Nonetheless, a method of tagging documents according to their intended helpfulness shall be suggested – one that draws on author goodwill in cases where authors wish to produce helpful texts that are easy to find – but otherwise this subject is left relatively unexplored for now. It is simply assumed here that it is in everyone's interest that users find exactly what they need as easily and quickly as possible. The last step is to actually modify search engines so that they may exploit the findings. This task is left to future work.

1.2 Defining the object under study

Search is a many-faceted kind of activity embracing various types and subtypes. The study of it involves a multitude of human processes associated with studies ranging from philosophy, perception, cognition and psychology over planning behaviour and device and interface interaction to reading and knowledge acquisition. To all this must of course be added the computational aspects, and one also has to take into account the classification and storage of knowledge (both mental and digital), as well as the collective (social) accumulation of knowledge in different domains. Finally, there are other social aspects such as recommendations by similar users, peer interaction, user classification and norm and preference recognition. This makes information retrieval research an exceptionally diverse and challenging field, but also one that is highly interesting and gratifying.

The very fact that these questions have been taken up enthusiastically by researchers from so many fields is both an asset and a challenge. Researchers from other fields often contribute valuable insights to our understanding, but at the same time difficulties in mutual understanding between researchers from different paradigms are frequent, and people may furthermore have very different agendas or foci, which tends to reduce clarity. This thesis attempts to bridge the gap between the strictly computational approaches and more humanistic methods and insights.

Research in information retrieval has mainly been conducted from three, originally different but currently converging, points of view[9]. Firstly, by librarians wishing to render their retrieval tools more effective at identifying relevant citations; secondly, by computer scientists seeking to improve web search; and thirdly, by lexicographers endeavouring to combine the insights from the two other camps with the aim of constructing advanced information tools. Most contributions can be ascribed to one of these traditions, even if they claim to have different aims from their predecessors.

Librarians have a long tradition of carrying out this kind of research. It is a basic idea of library cataloguing that it is possible in principle to provide a meaningful classification of resources that more or less matches the cognitive structuring of topics in users' minds, or at least one in which every document belongs in exactly one category (Ingwersen, 1992). This is by now widely acknowledged to be highly unrealistic; not only because the way people approach the search process is exceedingly varied, but also because of the very nature of information seeking; when people search for information, it is usually because they lack some crucial knowledge on the topic they are interested in, and consequently they may not know exactly what it covers and how it relates to other topics (Kuhlthau, 1993). Consequently, information seekers are rarely able to pinpoint exactly what they need. Instead, their needs develop through a process of what one tradition of library researchers refer to as *sensemaking* and *reduction of uncertainty* (Dervin & Dewdney, 1986). The realization of this led to fruitful progress in what has come to be called the *cognitive approach* to IR, of which the present contribution is an offshoot (Ingwersen, 1992; Ingwersen & Järvelin, 2005a).

Due to the notion of ideal classifications, bibliographic searches are performed on ordered datasets with topics manually added to every title by trained experts. This enables users to search for material

[9] Ingwersen (1992) provides an excellent review and discussion of the development of information science from its beginnings until the start of the 1990s, followed up in Ingwersen and Järvelin (2005a).

on a certain topic as assessed by an expert, but not necessarily to search the actual texts for subjects touched upon. At first, the contents of resources were accessible through title and topic labels only. However, as library materials are increasingly being digitalized the full text is now frequently searchable too. This has brought information retrieval researchers from the different camps together with a common goal so that the need for separate lines of research has diminished (Allan, Aslam, Belkin, & al., 2003).

With the advent of the internet, the need for effective search engines grew and a new line of research emerged out of the librarian tradition; one in which unordered datasets were searched (Saracevic, 1996). At first, the contents of these, too, were accessible through their titles only, but as with library materials, full documents came to be searchable later on. Contrary to library catalogues, however, web resources rarely contain satisfactory topic labels[10]. Also, while library tools were always expected to return references or links to titles, it is less clear what the entity of return is on the internet, since a site may contain very different pages, which may in turn consist of paragraphs with highly divergent meanings and purposes. Users would hardly be content to be led to the correct website – they wish to be shown to the exact place in the text where something relevant is found[11].

> (…) in general, text paragraphs and single figures or tables provide the smallest semantically confined retrieval passage of documents that may produce the information searched for. If we, for instance, observe how authors directly cite other text units in their own texts, such units are commonly paragraphs or sentences. (Ingwersen, 1996, p. 30)

Probably for this reason, the need for a much more subtle type of search engine was soon evident; one which could translate free text into topics on the fly and compare resource topics with queries. This was realized by dedicated search engine developers, but it never turned out perfect. There is still a long way to go before search engines are able to identify resources which are truly helpful for a particular user.

Lately, the idea of searching specifically for topics is well on its way to being abandoned entirely.

[10] It might reasonably be questioned whether library resources indeed carry satisfactory topic labels, i.e. whether this is at all possible. However, the topic labels have in this case at least been assigned by trained librarians, whereas labels in web resources are typically added by untrained authors and probably often without much thought as to their appropriateness.

[11] This point has recently been noted and a solution to it fairly successfully implemented by Google. It was originally investigated by Salton and associates in the 1990s under the name of *passage retrieval* (Salton, Allan, & Buckley, 1993).

> The idea that users want, or will be most happy with, documents on the topic of a search statement is an assertion that has absolutely no empirical support. (Harter, 1992, p. 603)

What emerges instead is a feeling that search engines should attempt to reconstruct the user's needs and identify documents that answer these needs; they should in fact be able to replace librarians (Ingwersen, 1992; Rose & Levinson, 2004). The 1980s saw a huge upsurge in the confidence in the powers of artificial intelligence, and some very ambitious projects were initiated which aimed at modelling the human mind and thereby not only replacing the librarian but the scientific experts themselves. In time, this was recognized to be an unrealistic project; human minds simply do not work like computers, and *vice versa* (Winograd & Flores, 1986). Except in very narrow domains where they sometimes proved useful, these ambitious projects were therefore abandoned. Instead, efforts were concentrated on producing artificial *mediators*, i.e. librarians, which would be able to understand the users' needs and help them find the correct resources (Ingwersen, 1992). However, a sufficient understanding of the linguistic issues was still lacking; queries continued to be short and ambiguous, and the only thing that one could do was make sure that as much as possible was known about the background of the user and the context of the search, in order that the needs might be reconstructed in spite of the deficient query. Some impressive user profile-based systems were designed for use in libraries and scientific databases, and the language issue was left for later work. A solution only came in the form of resignation by applying statistical methods instead of analysing and attending to precise wordings.

One of the problems with contextual systems that rely on the collection of large amounts of information on user and search context is the way such information is obtained. People simply do not want to spend time providing the necessary information, especially if they feel that they have already given all the information needed in their query. Sometimes, a few extra pieces of information on user needs can be collected without bothering the user too much. In the new millennium, lexicographers have truly entered the stage of information retrieval research with the goal of designing tools able to retrieve knowledge from an ordered dataset according to user needs. Traditionally, lexicographers were concerned with dictionaries and encyclopaedias only, but now these, too, are being digitalized, and, realizing that new talents are required, lexicographers have successfully taken up the challenge of designing digital tools (Bergenholtz & Gouws, 2010). In fact, the goals of lexicographers are virtually identical to those of IR researchers.

> [T]he essence of lexicography is its capacity to conceive and produce utility tools that can be consulted by specific types of users in order to acquire the type of information they may need in specific types of situations. (Tarp, 2008)

Building on the combined experience of librarians and search engine developers, lexicographers have designed new, advanced tools able to present data from an ordered dataset (known in lexicography as a *dictionary* in a broad sense) in different ways (called *visualizations*), depending on the user's needs.

What we need now is for search engine developers to take advantage of this new experience and develop tools for their own purposes which are able to take the user's needs into account in a similar way; not only encyclopaedia and dictionary users benefit from such approaches. Such work is under way, and various attempts have been made during the last couple of decades, with varying degrees of success. As the task is an extremely complex and difficult one, it is hardly surprising that it takes time to produce a satisfactory tool. At least now it is widely acknowledged that information retrieval requires an understanding of human needs, activities, motivations and intentions in order to be able to deliver truly useful results in response to user queries (Tamine-Lechani, Boughanem, & Daoud, 2010).

As already noted by Ingwersen (1992), the last 25 years have seen an unfortunate fragmentation of the research field, in addition to the three main schools, as new standpoints continually create new offshoots resulting in separate academic communities with little mutual interaction[12]. Quite apart from the obvious losses this creates for the involved researchers themselves, it has the unhappy effect that other academic fields do not perceive a unified field and therefore find it difficult to approach the subject, resulting in "a much smaller export ratio to related fields than deserved" (Ingwersen, 1992, p. VII). Ingwersen felt that the moment was ripe in the early 1990s, yet the situation is hardly any better today. On the contrary, multiple research centres have sprung up around the various commercial providers of information systems and retrieval tools, each with their own priorities and preferences as regards basic algorithms and user interfaces. And usability is not always given first priority.

1.2.1 Linguistically activated tools

First of all, we need to define the object that we are studying in order to delimit the research from some other types of IR research currently conducted. The present research is relevant to library tools, search engines and lexicographic tools alike. This begs the question whether other tools might likewise fit the bill. This is indeed the case, and in this section, a definition of the notion of **tool** for the purposes

[12] In a review of information science research, Pettigrew and McKechnie (2001) identified well over a hundred different theories originating from within the confines of the field and used in information science journals in the period 1993-1998.

of this thesis shall consequently be attempted. Intuitively, it is possible to classify tools according to at least the following criteria.

- The media employed and the kind of response expected from the tool: Responses may be either *informational* or *actional*, i.e. the tool may respond by retrieving information or by performing an action.

The present research is relevant for both kinds of tool, but as the informational one is by far the most common it will carry the main focus.

- The type and format of information searched: The data may be ordered as in a library or unordered as the web[13]. Furthermore, full-text or short descriptions only may be searchable.

The analyses apply equally to all of these types, but some of the methods proposed require extensive pre-indexing.

- The type and format of information provided by the user as part of the request: Interaction may be of a linguistic kind or consist in the mere pressing of a button or a hyperlink, etc. The tool might even be activated by the scanning of a picture, by iris recognition, etc.

In this thesis, only *linguistic interaction* will be considered, whether written or spoken (or in principle even signed). This is because of the interest in a correlation between a linguistic message and the intention behind it. It is also, incidentally, by far the commonest type of tool.

Thus, rather than confining the study to cases when users retrieve information with the help of a tool, we shall concern ourselves with those cases in which users employ language in communicating their needs to a tool, which may or may not respond in a way that contributes to the fulfilment of a task. This may properly be called a **linguistically activated tool** (or **LAT** for short).

There are good reasons for not restricting the present research to library tools, search engines proper or lexicographic tools; they all involve the use of language. Even other kinds of tools than these may be

[13] In fact, many web documents may be considered at least semi-ordered because they contain metadata and can be pre-indexed. Thus, the distinction between retrieval of ordered and unordered data has diminished (Allan, Aslam, Belkin, & al., 2003).

included to the extent that they may benefit from the results. Examples would be GPS devices, ticket machines, and museum information screens, to name but a few candidates[14].

A <u>linguistically activated tool</u> (LAT) is one to which users communicate their needs by way of language.

In practice, the focus will be on traditional search engines, since this is what most of the relevant literature is concerned with, and also where the empirical data come from. A **search engine** can be defined for our purposes as a LAT with an informational response. This excludes some tools which might reasonably have been included. It is, however, a practical definition which will prove useful to us at present.

A <u>search engine</u> is a linguistically activated tool designed to return informational responses.

Let us finally settle on a set of intuitive terms concerning the basic search situation. For simplicity, the term **search** shall be applied to the linguistic interaction with all LATs, even though some cases have very little in common with traditional searches. The **user** communicates his or her **needs** to the tool through a linguistic **query**, and if successful, the tool returns a **response**. An IR system being inherently a way of achieving some response based on the input, it is natural as a first attempt to classify queries according to the kind of response that the user is hoping (or expecting) to receive as a result of searching; responses may be **informational** or **actional**, and informational ones may be either **static** or **dynamic**. With a static response, a user locates information available in the document collection searched, whether or not its existence was known in advance. Dynamic responses, on the other hand, consist in the construction of new dynamic documents through online knowledge extraction from an ordered dataset; the user submits some information, which is used for returning information to the user. Dynamic responses imply a transfer of information in both directions. Databases do not contain ready-made knowledge; rather, knowledge emerges from the combination of data and query. On the other hand, static documents do contain knowledge ready for access, and the query merely serves to locate the information. Actional responses are actions performed in response to a query. Such responses produce changes in the physical world; the user transfers some information,

[14] Of course, these could be called search engines as well, given a different definition of that notion from the one adopted below. It is not our immediate objective to provide a definitive characterization of *search engines* here, but rather to explicate what kinds of *tool* are relevant to the present investigation.

but the response is non-informational, such as sending a product to the user or performing some action. By way of an example of static and dynamic responses, a search for *weather in Seattle* may generate static informational responses such as documents on the weather in that town in general, or else it may occasion a lookup in a meteorological database in order to present the user with actual, up-to-date meteorological data, i.e. a dynamic response.

An informational response may be delivered in the form of a **results list** containing references to the actual **resources** or **documents**, or simply by presenting them directly. Often, an excerpt called a **snippet** is shown with the link. Most of these core notions are explained more fully in due course. Here, it will suffice to point out that both of the terms *resource* and *document* will be taken to include all kinds of electronic files (Buckland, 1997), but whereas documents are restricted to "human work[s] with communicative intent directed towards other human beings and that is recorded in a material way" (Brier, 2008, p. 45)[15], resources include physical objects such as products bought from a webshop.

1.2.2 The role of a search engine – What to strive for

If we are to optimize search engines it is important to first agree on what it is that such tools should ideally achieve, and this is not as straightforward as it might initially appear. There are in fact two notions around of what constitutes the perfect response to a query. One is that it is a list of all relevant results ranked according to the degree of relevance so that no documents on the list are irrelevant and the most relevant ones are at the top.

Ideal 1: A search engine should return a ranked list of relevant results.

The other is that the perfect answer to a query is a set of documents that satisfy the needs of the user.

Ideal 2: A search engine should return a set of documents that satisfy the needs of the user.

Most researchers probably do not even think of the difference and thus have not formed a strict opinion on this matter, yet their definitions tend to reflect either one or the other of these viewpoints.

> The basic aim is to retrieve documents in order of their probability of relevance to the query (Belkin & Croft, 1987, p. 117)

[15] Brier's definition is in turn based on Buckland (1991a).

> The overall goal of IR [is] to ease the users' access to and the availability of information in documents (Ingwersen, 1992, p. 61)

The shift to the latter is associated with the rise of the cognitive approach, but the *ranked list* ideal has nevertheless survived to the present. Obviously, the two ideals are closely related but at the same time widely different. Strictly speaking, if the highest-ranking document answers all of the user's questions and needs, all other documents on the list should be irrelevant. Therefore, a list ought not to be very helpful. Perceived in this light, the *ranked list* option is not optimal. However, this objection rests on the assumption that the search engine is infallible and always places the best result at the top – which is of course entirely unrealistic – and also that the need is such that it can be satisfied by a single document. Kuhlthau (1993) also pointed out that *redundant* information (i.e., information that the user already has) is in fact essential for a successful search to take place. It is frequently taken for granted that information which is already known by the user is unnecessary and unwanted[16]. Indeed, it seems natural to think that an IR system should ideally not return information that the user is known to possess already; this is widely known as the *novelty* criterion[17]. Based on the findings of Kelly (1963), Kuhlthau (1993) dismisses both of these assumptions and reconstructs the process of converting new (and hence informative) information into known – and therefore subsequently *redundant* – information. She finds that the presence of redundant information is actually essential in order for a search to be successful, because we cannot acquire new information without incorporating it into what we already know. At the beginning of the search process, much of the encountered information on most topics is likely to be new, given that people do not usually search for information on topics of which they already have an intimate knowledge. This causes uncertainty and anxiety in the user, because new information must be incorporated into what is already known in order to make sense. There will presumably always be some redundant information in a document, so there is little risk that the new information cannot be incorporated at all (unless the text is highly technical and the user is a complete novice, or it is written in a language only partially mastered by the reader). Nonetheless, the amount of redundant information is decisive for the ease with which the user is able to acquire it. Redundant information is also reassuring and gives the user a feeling of being on the right track, and it is furthermore the user's existing body of knowledge which determines how the new information is interpreted and understood, and consequently also limits what is considered relevant. Consequently, it should not be dismissed as unwanted noise.

[16] See however Harter (1992) for an exception.
[17] See, among others, Barry and Schamber (1998), Choi and Rasmussen (2002), Maglaughlin and Sonnenwald (2002), Xu and Chen (2006) and Lopatovska and Mokros (2008).

> The balance of redundancy and uniqueness is critical in a search for information. Too much redundancy results in boredom; too much uniqueness causes anxiety. (Kuhlthau, 1993, p. 350)

Based on a study of librarians' use of search tools, Ingwersen (1982; 1992) concludes that the "provision of conceptual feedback from IR systems and the documents seems crucial for the user's request development and the match of participants' knowledge structures with the IR system" (Ingwersen, 1992, p. 101). Such an improvement of the ability to form a suitable query from the felt need would not be possible without some overlap of information between the possessed knowledge and that available from the resources shown. Thus, once again redundancy is shown to be important. Lack of redundancy may indeed cause users to abandon the search entirely in the early stages. As the process progresses, some of the unique (i.e. new) information is converted into redundant information; it has been understood and incorporated into the known. It is not entirely clear from Kuhlthau's description whether this happens the first time the information is read or whether only some of the information read is truly incorporated and hence made redundant. This is likely to depend on the user's interest in the topic, which, according to Kuhlthau's studies, typically grows during the search process as more is learnt about it. At any rate, the increasing amount of redundant information on the topic in question reassures the user and clarifies his or her thoughts. At a certain point, it becomes possible to perform a focused search for the exact piece of information needed. Contrary to what is frequently assumed to be the case, this was not possible at the earlier stages of the search because of the amount of uncertainty created by the new and unknown information.

One instance where the usefulness of redundant information is especially obvious is when users search for something they have already seen, and revisiting sites is in fact surprisingly frequent. Some studies show that between 60% and 80% of all pages seen have been visited earlier by the same user (McKenzie & Cockburn, 2001)[18]. One might also include in this category what has come to be known as *navigational* queries, i.e. queries intended to lead the user to a specific page by referring to its title (see chapter 7), since in such cases the user by definition either knows, assumes or at the very least hopes or imagines that a certain page exists. Even if the page has not been seen before and is not *known* to exist, the user effectively treats it as if it were indeed known in advance. As suggested in section 7.3, refinding information is a different psychological process from finding something unknown.

[18] Note, however, that this figure unfortunately includes pages that may not have the same content every time it is accessed, such as news portals and social sites, which of course does not qualify as refinding the same information.

We shall mostly be interested in searches where users are looking for something that is not known in advance.

One further assumption, which – perhaps surprisingly – turns out to be unfounded, is the one that users always prefer the most highly relevant items when submitting a query to an IR system. This has been shown by Spink *et al.* (1998) to be a mistaken belief, in particular as regards users who are in the initial phases of a search process (cf. section 2.1.3). In other words, a ranked list may be an appropriate compromise if not an ideal.

In the end, what users need is the perfect collection of documents containing the right information (also taking into account what their preferred sources are, etc.[19]), couched in a setting that allows them to make optimal use of it. In other words, the information should be easily accessible, both logistically and when it comes to the cost (in time, money, efforts and emotional strain) of retrieving the resources, and furthermore texts should be written in a style easily understood by the user, should not presuppose background knowledge not possessed by the user, and should be easily applicable to the task at hand. These are strong criteria indeed, and they are rarely met in full. However, we must be clear about the ideal situation in order not to aim in the wrong direction.

> Ideal 3: A search engine should present the user with any available collection of resources that answer the user's information needs while at the same time (ideally) being highly accessible. The information contained in the documents should also (ideally) be highly accessible and easily applicable to the task at hand.

Such resources may well be presented to the user in the form of a ranked list, but ranking is crucially not the ultimate *purpose* of the tool; it is merely a by-product or a means to an end[20]. In fact, different kinds of ranking which do not simply grade resources according to presumed relevance but encourage user interaction in order to identify the ultimate need may in many cases be preferable (Brandt, Joachims, Yue, & Bank, 2011).

The *ranked list* definition dates back to the days when library tools were merely expected to provide a list of references without access to the original documents. In this context it works, because if you do

[19] Users may prefer to get the information from certain sources rather than others, or they may want to ensure that the information is widely accepted.

[20] It should be mentioned in passing that some types of search actually do end with the presentation of results, such as when people search for a term simply in order to find out whether it is correctly spelled, or how commonly it occurs, and so on. Such special uses, though actually very common these days, are of less importance to us here since they are usually not the main function of the tool, although they might be.

not have direct access to the actual document you cannot know whether it will actually satisfy your needs, so you will prefer to have a list of potential resources rather than only receiving the one reference which the system considers most relevant. Perhaps unfortunately, it has outlived the introduction of full text search and has remained a widespread interpretation of the role of a search engine.

> The goal of personalized search systems is to provide differently ordered results for similar or identical queries when issued by different users with different information needs. (Lee, Kim, & Lee, 2010, p. 76)

This view can easily be proved not to be very useful as an ideal since we now know that information needs change as users read the results list, thus immediately making the list obsolete (Harter, 1992). As noted by Bates (1989, p. 410), "the query is satisfied not by a single final retrieved set, but by a series of selections of individual references and bits of information at each stage of the ever-modifying search". In conclusion, then, it would seem that the best we can do for our users is provide them with some sort of interface (e.g. a ranked results list or a dynamic interactive interface, etc.) that will enable them to pick those documents which are most likely to prove helpful to them.

Harter (1992) not only endorses the traditional view; he even suggests that the cognitive changes caused by reading the results list are what really matters, rather than those resulting from reading the actual documents, as these "are in some sense less important than citations that themselves cause cognitive change directly" (Harter, 1992, p. 612). This shows us that he is taking the standard point of view of a librarian; he is interested in "references that allow the making of new intellectual connections or cause other cognitive change" because these references are exactly those that will produce even more references. However, surely the information provided by relevant documents is more important to users of a LAT than mere references to them, even when these do cause people to think in new and creative ways? Some researchers interested in personalized search have adopted the latter view.

> The main goal of an information retrieval system is to return to the user the most valuable documents in response to his queries. (Tamine-Lechani, Boughanem, & Daoud, 2010, p. 2)

> The goal of a search engine is to have indexed the right pages for any user search, and to effectively retrieve them in response to the query supplied by the user. (Grimes, Tang, & Russell, 2007)

Compare also the following carefully stated definition given in Mariam Daoud's doctoral thesis.

> Un système de recherche d'information (RI) est un système qui permet de retrouver, à partir d'une collection de documents, les documents susceptibles d'être pertinents à un besoin en information d'un utilisateur exprimé sous forme d'une requête.
>
> 'An information retrieval (IR) system is a system that allows you to find, from a collection of documents, those documents which are likely to be pertinent with regard to an information need of a user, expressed in the form of a query' (Daoud M. , 2009, p. 30; my translation)

What is lacking in this description of the ideal system is a definition of what it means for something to be *pertinent*, and also what query expression caused the *likelihood* in question. These are the contributions that we shall attempt to make in Part II and III, respectively.

The two competing views on what the task of a search engine is are probably here to stay; librarians and others like them will maintain that its function is to retrieve relevant citations, while probably most others (including users) will find it more helpful for a search engine to identify relevant and useful documents, or as Kalervo Järvelin (2011) put it, "The ultimate goal of information retrieval (IR) research is to create ways to support humans to better access information in order to better carry out their (work) tasks". There is no doubt that a ranked results list is a useful means of getting there, so the difference is mainly in the conceptualization of the problem. It may be considered an unimportant, academic question, yet some of the tools covered by the present thesis according to the definition above would not have a results list at all, for the simple reason that they would not have informational responses. Also, the distinction is crucial when it comes to evaluating the effectiveness of different solutions to the retrieval problem, something which is considered of great importance as it lends an air of scientific accuracy to the study of IR.

1.2.3 Enterprise search and log analysis

Enterprise search research deals not with web search but with search systems developed for use on smaller resource collections such as individual home pages or information-rich sites, or intrawebs. While library search systems and lexicographic systems are very much focused on the retrieval of information and (more particularly) knowledge, the systems that we shall be interested in here have a much wider range of possible intentions associated with them. Among many other tasks, users of these systems may want to buy a kind of product, find a suitable job, have the system perform some action, receive help with an activity, locate particular facts, obtain definitions of certain concepts, and of course find information on certain topics and extend their knowledge and improve their understanding of more or less specific topics. The overall diversity to be dealt with by enterprise search research as a whole is thus just as overwhelming as with web search, but the individual site or intraweb has a much more limited range of uses. The precise range varies greatly between systems, as some sites combine several of these characteristics, e.g. product sales, corporate information and support facilities and services (cf. the TDC homepage used for the empirical analyses below), or research papers, manuals and directives, personal information concerning staff, etc. (as in many intrawebs).

Log analysis as a means to studying information retrieval interaction is first and foremost considered a way of subjecting large amounts of pseudo-behavioural data to statistical analysis.

> The major limitations of these studies include that they only catch a narrow facet of the searcher's Web interaction. The searcher, his/her intentionality, strategies, and motivations are hardly known. On the other hand, log analysis is an easy way of getting hold of data that can be treated with quantitative methods. We can use the studies to obtain statistically significant data about searchers' selection of search keys and use of syntax in queries (Ingwersen & Järvelin, 2005a, p. 225)

It certainly is the superior method for such studies. For more interesting data, behavioural studies need to be carried out. However, the query is almost the only data available to the system run-time, apart from the occasional contextual clue, and so, from a search engine developer's point of view, this is the kind of information that we need most urgently to be able to interpret as effectively as possible.

Is it not possible that with the right methods to hand we will be able to get more information out of queries in future?

With the models and methods developed in this thesis, a possibility is indeed opened up of allowing studies of user intentions and states of mind through log analysis, which has not hitherto been considered possible due to the nature of queries (Pharo & Järvelin, 2004).

Because queries consist of pieces of language, they do inherently convey intentions and motivations, even if the clues are rather scanty to the extent that the queries are removed from their original context. However, enterprise search allows us to reconstruct that context to a much wider extent than web search, and so by focusing on the expressive choices made by searchers and remaining aware of the inherent ambiguities, it is possible to extract a reasonable amount of highly interesting information on the users of a site. What users choose to refer to with their queries reveals much about the way they approach a task and the cognitive state they are in, such as their level of uncertainty and anxiety about the search task. At the same time, there is no doubt that, when possible, combining such studies with observation of user behaviour provides even more detailed information and may help eliminate some of the ambiguity.

1.2.4 Constructing linguistically activated tools

As pointed out in the youth of artificial intelligence by Winograd and Flores (1986, p. 52f), there is an inherent paradox in building intelligent and user-friendly computer systems. The simplest computer programs perform specific actions in a carefully stated sequence, but as systems become more complex and adapted to human needs, they are gradually made more and more flexible by focusing on goals and potential actions rather than scripts. This enables the system to react to changes and unforeseen situations. However, the entire enterprise hinges on the ability of programmers to predict to what uses the program will be put, and this is not as straightforward as one might think. A good word processor or other similar application allows the user to take advantage of a broad range of tools as the needs arise because such an application will invariably be used in a large number of ways that the designers did not and could not anticipate. The same is true of search engines and many other LATs which can be used in an endless number of ways. It is effectively impossible to anticipate what kinds of requests will be made to a search engine, even within a narrow domain such as the job search system analysed in this thesis.

> The most successful designs are not those that try to fully model the domain in which they operate, but those that are 'in alignment' with the fundamental structure of that domain, and that allow for modification and evolution to generate new structural coupling. As observers (and programmers), we want to understand to the best of our ability just what the relevant domain of action is. This understanding guides our design and selection of structural changes, but need not (and in fact cannot) be embodied in the form of the mechanism. (Winograd & Flores, 1986, p. 53)

However, some uses are always more common or likely than others, and so it is nevertheless possible to define a set of uses that must be supported by a certain LAT. Users with other, less common needs will then not be specifically catered for, but must manage as best they can. One way of helping such users is by providing additional tools, as in the word processor case, such as various filters and views and other options. These are repeatedly reported not to be used to any significant degree in studies of IR user behaviour (Jacsó, 2006; Azzopardi, 2009). However, such claims completely miss the point that the additional features are specifically created for the few with special needs. Hence, it is not to be expected that any one functionality be used on a regular basis. On the contrary, if they were, that would be a sign that the system was badly tuned to the predominant user types. As there will always be some users with more particular needs, it is important to keep such tools within reach, while at the same time making sure that they are only needed in exceptional cases. It is the job of the system designer to create a system that limits the need to apply special functionalities as much as possible, seeing as there is no doubt that the user studies are right in claiming that users prefer not to bother about settings and options but simply want an immediate, relevant response to their vague and ambiguous query.

Crucially, the choice of what functionalities to support should be guided not by the rationalistic approach that people tend to take but by a behavioural one supported by cognitive insights (Winograd & Flores, 1986, p. 71f). This is because, contrary to popular belief, most human behaviour is *not* rationalistic but is driven by often subconscious choices made in a domain of potential actions in response to problematic situations, or *breakdowns*, to use a term from Martin Heidegger's philosophy (Heidegger, 1996). A breakdown is a situation in which a person hits a stumbling block and suddenly becomes aware of the need to be conscious about the various possibilities available for progressing further. This is when users start thinking about how to use tools most effectively, and what actions they should take. It does not follow, however, that they are conscious of the reasoning behind the actual choices that are made in this space of possibilities, or that the possible actions and outcomes are clearly distinguished and delimited in the person's mind. Rationality is only a minor ingredient in the process.

Thus, far from agreeing with Barry (1994, p. 158) that "before there can be any discussion of changing retrieval mechanisms to include user-defined criteria, we should determine whether users can actually identify those criteria as part of the search request", I would urge designers to concentrate on actual user behaviour rather than conscious reports about preferences; relevance criteria, for instance, may well be present in some form without the user being any the wiser.

1.3 Improving enterprise search

The system designer has the choice between 1) requiring the user to master a specialized search language designed to supply the system with the exact information needed in order to be able to retrieve what the user wants, 2) letting users express their needs in any way they want, using more or less natural language (though as we shall see this often involves a specialized usage restricted to search interaction, but developed as a natural consequence of interacting with such systems), or 3) doing the latter, but supplementing queries with additional information on the users and their needs by interrogating them or supplying them with forms and menus etc. The first alternative is only feasible if the system is tailored for use by experts who have received training in the search language. In other cases, allowing natural language queries is the only real option, with or without additional input. It turns out that users are not often willing to supply the additional information – especially not before executing the search (Jacsó, 2006). Thus, as a system designer, one should be prepared to accept that the queries are the only input the system is going to get in the majority of cases. The exception is contextual information which is available without the conscious intervention of the user, such as geographic location (especially in the case of portable devices) and the type of site consulted. Also, some information can be derived from user behaviour, such as click behaviour and dwell times.

From the designer's point of view, the usefulness of search results can logically be improved in two ways. Either a different (or at least differently-ranked) set of documents is retrieved in response to the user's original query in order to better match the user's presumed needs, or the query is manipulated in some way or other, by methods known under the umbrella term of **query expansion**, leading in turn to a different outcome (Bhogal, Macfarlane, & Smith, 2007). Furthermore, any amendment of result retrieval may either be based on the assumed needs of typical users of the tool in question, or, preferably, the actual needs of the current user. Building a satisfactory model of the user – a **user profile** – is no mean feat (Luxenburger, 2008; Daoud M. , 2009). A variety of different solutions for the detection of user **intentions** have been devised during the past couple of decades. The investigations centre on how to define and access the relevant contextual parameters that contribute to the difference in interpretation from one query to the next.

1.3.1 A brief history of search research

IR research was born in the 1940s, but only really set off with the famous Cranfield studies in the 1960s. The kind of search performed at that time was Boolean and involved simple matching of queries to document strings. Since then, tremendous progress has been made in the technologies used and the understanding of the problems involved. Critique of the existing systems grew, due to their focus on algorithms and consequent neglect of user behaviour and psychology, so during the early 1980s this traditional *system-oriented* approach to IR was replaced by a *user-oriented* type, the idea being that a search engine should find what the user needed, rather than what matched a given string. This meant that results were no longer objectively definable as either relevant or irrelevant; now that depended on the user's intentions and goals with the search. A user-oriented system is built around a user profile that should ideally record what the user is interested in, what state he or she is in at the start of the process, what abilities the user possesses, etc.

The labels that name various approaches to IR are partly defined on the basis of differences between the systems themselves, but more often either on the way the problem is conceptualized by the designers, or the way they go about their research and test their systems. The names thus primarily refer to research paradigms and only secondarily to differences in the actual system types. The difference between system-oriented and user-oriented types is very much one of conceptualization, but various offshoots of user-oriented IR were to arise during the following years. *Interactive IR* is one such type that is mainly defined by the way user behaviour is investigated in order to study how systems should be designed.

> IR interaction (…) stresses the iterative nature of information searching. That is, it seeks to better understand phenomena such as search strategies, search term generation and use, and successive searches by users for the same research problem. Methods employed to study these phenomena include observation of users in naturalistic settings, discourse analysis, and other protocol analyses such as think aloud protocols. (Robins, 2000, p. 58)

With the advent of AI methods, user-oriented IR grew into a *cognitive* approach on the conceptualization side that took into account what had been learnt about the workings of the human mind in that branch of science. Consequently, new sophisticated systems were developed on the basis of intricate models of human behaviour and need development. Many of these systems were able to actively seek the information from the user that they needed in order to help them come to grips with their true needs. What such systems rarely bothered with was really understanding the natural language strings that made up the queries. Indeed, this was generally thought not only impossible but also unnecessary.

> (…) if an intermediary mechanism translates a request it will be on the premises of its knowledge-base, which either is implemented by another individual, mirroring his conceptual structures, goals, etc., or it is acquired by some rules via NL processing of texts and/or user statements. Consequently, the translated meaning will tend to be any other meaning than that of the actual user. (Ingwersen, 1992, p. 192)

However, this problem was to some extent diminished with the advent of contextualization methods. Ingwersen himself foresaw the development of a new *contextual* approach to IR, and it soon became the leading branch of research. By gathering information on the context ambiguous terms could be disambiguated and needs could be specified more precisely without requiring the user to supply the facts. Incidentally, this opened the door to understanding queries more properly, since context by definition includes what is decisive for the proper interpretation of the message. If a query cannot be interpreted properly because the user's knowledge base or conceptual structures differ from those of the system, then this is because the user subconsciously expects the system to share such structures. After all, if this were not the case, there could be no hope of performing successful communication. The idea is then to collect information on the user's expectations and beliefs, thereby allowing the system to interpret the query as intended from the user's point of view. This is not to say, of course, that such an endeavour is in any way easy in practice; quite on the contrary.

Originally, a search system was just a tool used by experts as part of a more encompassing process of locating information. Much of the actual work was done by the users, because they had to formulate their needs while anticipating the workings of the system and guess how to retrieve useful information that would be applicable to their task. With the user-oriented approach, and especially with cognitive IR, the optimism grew of actually modelling this piece of human behaviour within the tool so that the user would be relieved of that work and could concentrate on the main work task. Ingwersen (1992) criticized these *cognitivistic* attempts for being unrealistic and unnecessarily optimistic.

> As the author sees the present IR research situation, there exist[s] a danger of falling into the 'cognitivistic' trap, i.e. to begin to believe that the inference mechanism may actually provide all the necessary 'knowledge' or information processing for successful retrieval, based on expert system-like question-answering during model building.
>
> This cognitivistic AI solution is a dead-end for IR, and one may view this research position as a left-over from the traditional research approach. (Ingwersen, 1992, p. 180f)

In other words, the system should not pretend to be a model of the human brain itself; computers do not work like brains and real humans always perform these tasks in a superior way. Thus, the standard method in the new millennium became a more moderated version of cognitive IR informed by AI and cognitive science but not attempting to do the actual thinking for the user. The system attempts to locate useful information based on information provided by the user explicitly and implicitly through interactive behaviour and drawn from context, but the user is still required to look through the results returned and select those which seem to be most useful in the situation at hand. Today, the latest trends seem to be a further retreat from the optimistic barricades of cognitivism by relying less and less on the system's ability to predict usefulness and leaving it to the user to decide instead. Obviously, with the constantly growing amounts of documents, it is still a major task of search engines to sort through the masses and identify what might potentially be relevant, but the tendency is to nevertheless present the user with a choice. This can be done by making sure that the first page of results is maximally diverse, both as regards the interpretation of the query terms and the types of resources (Wang & Zhu, 2009). Another promising development is *dynamically ranked IR*, which builds on the diversity concept by adding an interactive functionality that allows the user to respond to the suggestions and click a result which seems promising in order to access other similar results (Brandt, Joachims, Yue, & Bank, 2011). This in effect concentrates the results of a search on the first results page by condensing them to a set of distinct types. Perhaps this will in turn lead to a new paradigm of **dynamic IR** to complement the others. Here, we shall assign all attempts at coping with diversity and ambiguity that rely on interface and user interactivity such as presenting maximally diversified results or allowing users to manipulate the presentation interactively to this dynamic approach which does not seem to have received an established name for itself previously.

1.3.2 What further research can teach us

Seeing the immense success of web search systems such as the one developed by Google during recent years, it might reasonably be asked what can be gained by continued research into information retrieval.

Is it really worth the effort trying to improve something that is already so effective?

Many of the problems explored in this thesis have been treated previously by computer scientists, and sometimes with very successful results. Highly efficient search engines have been built, which are able to handle an astonishing amount of data in a fraction of a second and even to incorporate some contextual information through user profiles and the utilization of different kinds of feedback. These systems typically deploy a minimum of linguistic knowledge and instead rely almost entirely on statistical data extracted from the huge corpus that is the World Wide Web. So why is it that a linguistic approach is in any way desirable, if computer scientists do such a marvellous job of it?

As pointed out by Allan *et al.* (2003), there are several reasons for retaining an interest in the topic. Firstly, information retrieval is simply much more than just web search. There is a wide range of information systems that make use of information retrieval techniques in different ways, and they all benefit from different methods. It is apparent from the success of Google and others that statistical methods are highly efficient when it comes to web search[21]. This is not the case in narrower domains such as searching an intranet or the documents available from a certain homepage. Statistical methods require huge amounts of data in order to work and are accordingly suited only to web search and not to search in more restricted corpora such as enterprise or intranet search. In more limited collections, other types of system are more effective, simply because statistics are less reliable when used on a smaller sample. Also, even though one might think that statistical results might be achieved on the large corpus and then applied to a smaller domain, the kinds of queries which are submitted to web search engines are not representative of all kinds of queries. In different kinds of system, users have different types of needs and will consequently not be helped by the same means. General information is simply not the same as domain-specific information, and so different, more precise systems are needed when all documents pertain to a somewhat narrow domain. Also, user behaviour is affected by the awareness of the larger size of the web document collection (Blair, 2002a). Finally, current web search algorithms combine similarity and popularity ranking in order to get rid of the large amounts of less useful resources, which is not necessarily advantageous in an enterprise search environment where every document may be important.

> [It is] imperative to treat large-scale systems as fundamentally different from small-scale systems. That is, large-scale document retrieval systems may require not only a different design model, but may also require a substantially different theoretical foundation than small systems. (Blair, 2002a, p. 282)

Web search will never be perfect, because – apart from the many other obstacles – it is required to be suitable to all kinds of users searching for documents across all domains (Teevan, Alvarado, Ackerman, & Karger, 2004). Consequently, statistics is an appropriate choice. Even so, as pointed out by Allan *et al.* (2003) in a report from a workshop on the short- and long-term challenges for future IR research the generalizing tendency may be detrimental even to web search now that we have got this far.

> Much active research in information retrieval is carried out by abstracting the user away from the problem: judgments are captured and held constant, non-binary relevance is ignored as too complex, evaluation is on a single round of query-results without opportunity to adjust the query, and so on. This abstraction has been incredibly successful in enabling research to advance rapidly, creating more effective and efficient systems for retrieving and organizing information. However, those

[21] However, de Lima & Pedersen (1999) report evidence of linguistic analysis outperforming statistical methods even in a web search context.

> improvements in retrieval accuracy appear to have dwindled in the past half dozen years. It may be that one reason researchers are unable to advance beyond the current plateau is that the evaluation model forces systems toward user-generic approaches that are "good enough" for everyone, and therefore "never great" for anyone. (Allan, Aslam, Belkin, & al., 2003, p. 38)

User behaviour is intimately connected with the size of the resource collection. People adapt to the requirements of a given system because they have to; searching for a highly ambiguous term in a web environment is simply not feasible without providing some disambiguating material as well. Acknowledging such system-aware behaviour, three basic functions of queries have been identified by various authors, though usually treated in pairs: Firstly, some query terms are intended for *description* and others for *discrimination* (Blair, 2003; Terney, 2009). The former class of terms describe either the problem state or the desired resources. Discrimination terms, on the other hand, exist to filter away unwanted irrelevant resources that are known or assumed to exist and to potentially interfere with the search. For instance, a semantically disambiguating term such as *car* in the imaginary query *jaguar car* is likely to be added as a mere discriminator so that only documents on cars, rather than felines, are returned. It is not necessary that the word *car* be included in the document as such if the system is able to disambiguate the sense by semantic analysis. We shall call describing and discriminating terms **descriptors** and **discriminators**, respectively, and we shall say that discriminators require **meta-awareness**, i.e. awareness of the requirements of the system in addition to conscience of the need to communicate what is desired. Only when such awareness is triggered – typically through an initial failure to retrieve the wanted resources – can users supply discriminators that filter out the unwanted noise recognized to be available in the collection and to potentially interfere with the search. Blair (2003) predicts that discriminators will become increasingly frequent in queries as resource collections continue to grow.

> (…) as document collections grow larger and larger, a subtle change in the information retrieval process is taking place. Instead of the goal of search query formulation being primarily the description of what is wanted, the overriding goal of query formulation has become the discrimination of small numbers of desirable documents from increasingly large numbers of unwanted documents. (Blair, 2003, p. 8)

The amount of irrelevant material is simply bound to be increasing faster than the number of relevant resources.

> (…) the increase in the size of the document retrieval system may not be increasing the number of documents relevant to the searcher's request by very much, it may only be increasing the size of the retrieved set the searcher must browse through to find what he wants – the haystack in which the searcher must find the needle is simply getting larger. (Blair, 2002a, p. 279)

This follows naturally from the observation by Zipf (1949) that the words contained in documents will get increasingly semantically ambiguous with a larger collection size. Blair suggests that we take the consequence of this and develop methods based on what he calls *partitioning search*. Such a system would start by filtering out what it "believes" to be clearly irrelevant material, thus in effect achieving a state in which only a reasonably small collection is searched in order to find the truly *relevant* resources. This in turn would allow the subsequent application of methods such as those developed here.

For domain search, one should definitely take advantage of the more predictable context and help users get more precise results. Additionally, it is probable that people are on average willing to put more effort into searching on a particular site in order to find exactly what is needed (Balog, Weerkamp, & de Rijke, 2008). They are also likely to have higher expectations of the quality of results if they are searching a single site than when they try their luck on the entire web. In such cases, it is desirable that a search tool be able to decode the query more literally, interpreting it as an instance of communication from the user to the system rather than as a string to be matched with a certain degree of statistical probability. Consequently, more effort should be put into perfecting local search systems. As pointed out by Allan *et al.* (2003, p. 47), "the approaches and experimental resources that brought the field to its current level of success will not be sufficient to reach the next level". In other words, far from being an exhausted subject, IR research is as much in need of innovative thinking as ever.

1.3.3 Structure of the thesis

It is clear even from the above short introduction that there are a great many problems relating to the task ahead. On the theoretical side alone, *relevance*, *context*, *information need*, *intention*, *information*, *topic* and *reference* are among the controversial subjects to be dealt with one at a time. They are all intimately connected, and it is impossible to isolate one without taking the others into account if we are to understand how users communicate their needs to a search tool.

For clarity and ease of digestion, the thesis is organized in three separate parts, each with its own partial goals that contribute to the overall understanding of query expression. The thesis starts with some highly theoretical and abstract discussions, but gradually takes the reader towards more concrete issues of query expression as it progresses.

Part I looks at some of the cognitive and psychological aspects of search from a theoretical point of view by exploring the way user needs develop and the role they play in query formulation. The **Faceted Stimulus** model is introduced to explain some of the complexities of need development and relevance assessment. Furthermore, attention is paid to the contextual inferences required for interpreting the intention behind a query. While it is easy to dismiss variation in query form as being merely due to "contextual effects", context is in fact an integral part of the search process and not just an environment in which it takes place (Kelly D. , 2006). Even so, some aspects of it are clearly more important than others when it comes to the formulation of a query. A model of context is suggested based on the insights provided by the Faceted Stimulus model. Finally, the dataset used for the empirical analyses is presented.

Part II is a study of what makes a document relevant. Clearly, if we are to help users get more relevant results, it is vital that we understand this in order that we may isolate the effects of relevance judgment from those of need recognition in the query construction process. The phenomenon of relevance is at the heart of information retrieval, but even so no unifying model of relevance has as yet been forthcoming, and there is not even a satisfactory definition of the concept. Here, a new overarching model by the name of **Regulated Flux** is introduced which explains the relative fluidity of relevance assessment tests while at the same time providing the limits that regulate the process. Also, a definition and basic understanding of relevance is suggested.

Based on the previous parts, **Part III** dives into the question of what queries refer to and the development of a model to explain query variation as a result of different choices during the search process. Also explored is the kind of variation that occurs in query formulation, especially in the way of morphological and syntactic variation. It is interesting to contemplate what syntactic and morphological rules might have evolved as regards the construction of queries, and the empirical data support the supposition that queries do indeed tend to adhere to such rules. The suggested **Stage-based Intent** model of query expression choices and search behaviour fits beautifully into the Faceted Stimulus and Regulated Flux models and explains the residual variation in query expression found in the empirical studies. Finally, a method for analysing search logs with reference to the expression alone is suggested by the name of **Target Analysis**.

The theoretical discussions are backed with log material from the two log files L^{TDC} and L^{WID} introduced in chapter 3 below. Some extensive studies of these logs are included in Part III, while several smaller studies are sprinkled throughout the chapters for illustration of the issues treated.

The main purpose of these various steps is to produce one general picture where all the elements can be seen to have their natural part to play. Thus, the thesis is rounded off by combining the models into one overarching model of the search process that includes all of the studied sub-processes. In many cases, the suggested models do not enable us to accurately predict user behaviour in a given situation, but they do at least explain the choices that users make. Further research will be needed in order to get closer to a stage of research at which choices can be predicted to any extent in order that it may be incorporated into working algorithms. At this point, simply acquiring an understanding of the issues involved is crucial. Without such understanding, any attempt to incorporate user behaviour and relevance assessment into the IR process must be spurious. Many have tried through the years, and some progress has definitely been made, but always there are the same stumbling blocks; what does it mean for something to be relevant, and how does a query relate to what the user needs? It is my hope that the present work will make these issues just a little clearer to the reader.

2 Search as a process

> *The language of retrieval not only limits how we articulate what we want but can also constrain the very thought process in which we determine what we want. Presumably, we would like to think that we mold our information retrieval systems to serve our need for finding information; but, if Wittgenstein is correct, then it may be the case that our information retrieval systems are molding us to think along their lines. If this is the case, then it may be extremely difficult to design radically different or improved retrieval systems, because we are virtually locked into the way of thinking about retrieval that is embodied by existing systems.*
>
> (Blair, 2003)

Expressing your needs through a query is an act of communication from a user to a system. It is therefore natural to assume that queries are the counterparts of questions in discourse, as both are requests for information.

> Information retrieval systems, expert systems, management and decision information systems, reference services and so on, are instituted to answer questions by users – this is their reason for existence and their basic objective, and this is (or at least should be) the overriding feature in their design. (Saracevic, Kantor, Chamis, & Trivison, 1988)[22]

[22] Note that the questions treated by these authors really are full questions put to an intermediary. Nowadays, such questions are not part of standard IR interaction. Their notion of query is also highly non-standard, at least today: "The query is the specific aspect asked about the question's subject. It is an attribute, characteristic, component, or part of the subject about which information is desired" (Saracevic, Kantor, Chamis, & Trivison, 1988, p. 166). The example they give is this: (What are)$_{\text{Lead-in}}$ (the advertising expenditures)$_{\text{query}}$ (of the automobile industry)$_{\text{subject}}$.

This is such a commonly made assumption that it is even reflected in the widespread use of the term **request** when speaking of search, and indeed in the word **query**, both of which terms ultimately derive from Latin *quaerō*, 'I ask' (see box *Query*). As with so much in IR, this assumption is a heritage from the librarians that laid its foundations. In the library interview, users do indeed frequently state their needs in the form of questions to the librarian, and this has been incorporated into the literature on information needs from early on, to the point of being considered an integral part of a need (Taylor R. S., 1962). In spite of this honourable pedigree, there are compelling reasons not to view queries as questions but rather as a special form of requests more akin to imperatives than interrogatives (Sparck Jones, 1989). For instance, a user who wants to know what the capital of Germany is might submit the query *the capital of Germany*. This is a fact-finding query – the user is looking for a simple piece of information answering the needs that seem to represent the informational gap in his or her mind. In discourse, this would be expressed as *What is the capital of Germany?*, where the information gap is signalled with the question word *what*. When the interlocutor is a machine, however, people typically reduce this utterance to *the capital of Germany*, or even just *Germany*, i.e. it generally ceases to be a question and becomes a request for information on a certain topic. In other words, people may want answers to questions or information on topics, but when interacting with LATs they tend to *express* their needs in terms of the latter, whereas in discourse the reverse is true[23].

Query

query *n.* question, inquiry. Before 1635 *quaery*, alteration (influenced by inquiry) of earlier *quere, quaere* question (1589); borrowed from Latin *quaere* ask, imperative of *quaerere* to seek, gain, ask, of unknown origin. The spelling *query* is first recorded in 1645.

(Chambers Dictionary of Etymology, 1988)

Questions are often thought of as a kind of gap-filler; the speaker has a gap in his or her knowledge and produces a question that frames this need as something to be filled by the addressee. This idea is typically brought over to the understanding of queries in information retrieval (Dervin & Nilan, 1986). It, too, probably has its roots in the library tradition. The *sensemaking* theory was originally developed for information search in the library context (and hence with librarians as the "system"[24], rather than computers), but was later introduced to IR research in general (Kuhlthau, 1993; Dervin B. , 2003). It utilizes the so-called *situations-gaps-uses model* in which the *situation* refers to "those events in a

[23] However, log analysis shows that questions are in fact also surprisingly common in search, cf. section 9.3.2.
[24] According to Taylor (1968), library users really do consider the librarian to be part of the system at the point when they start developing their formalized need (cf. section 2.2.1). Unlike at the earlier stages, the librarian is then thought to organize his or her categories in the same formalized way as the library shelves themselves, which represent the actual system.

person's life that create the context for a lack of sense, or a gap, i.e., an occurrence that raises questions", the *gap* "is translated into question form during the reference interview, and the answer to the question may be seen as a bridge across the gap", and the *use* is what the user hopes to be able to do after crossing the bridge (Dervin & Dewdney, 1986, p. 2). Even where there is no talk of questions, however, search engines are still often thought of as filling an information gap.

> Usually a query is an expression of a lack of information, a retrieved document is intended to fill that information gap. (van Rijsbergen, 1990, p. 35)

> [An information need is] a consciously identified gap in the knowledge available to [a user]. (Ingwersen & Järvelin, 2005a, p. 20)

Thus, Tamine-Lechani *et al.* suggest that the effectiveness of a user-oriented information retrieval system (as opposed to a traditional system-oriented one) be measured as its ability to "satisfy the information gap in a specific search context" (2010, p. 4). The problem is that this implies the existence of a mass of knowledge in which the gap is to be found. Indeed, how can you satisfy an information gap if it is not acknowledged beforehand? However, if a user were to search for *Berlin*, it would probably not be because this person did not know anything about Berlin, and so *Berlin* itself does not represent a gap. Nor is it probable that he or she was thinking of any specific question concerning the German capital. Rather, it is a request to *TELL ME SOMETHING/ANYTHING/EVERYTHING ABOUT BERLIN*. Since much of the response information is likely not to have been foreseen by the user, it does not fill an acknowledged gap, but rather builds the very structure of knowledge in which gaps may then be identified along the way.

Let us look at the following example from one of the search logs investigated below.

2.1	**landenummer 0037**	'country code 0037'	TDC (1)

This query, submitted to the telecommunications company TDC, suggests a question such as *WHAT COUNTRY HAS THE COUNTRY CODE 0037?* Nonetheless, due to the so-called *label effect* (Ingwersen, 1982), it has not been expressed as a question but as an informational entity that the user wants clarification about, i.e. *TELL ME ABOUT COUNTRY CODE 0037.* The whole idea of the query as representing a gap is a remnant of the notion that queries are questions. However tempting the idea might be, it must be emphasized that most queries are not reductions of questions in the traditional sense. Rather, they are prototypically keywords referring to topics on which the user wants to increase his or her knowledge and references to problematic or desired situations, etc. There is in fact a surprising amount of variability in this regard, as we shall see in Part III. Some people do pose questions to the LAT, some present labels or descriptions of situations, and some give direct commands in order to tell the LAT what they require. The best description of a query is probably that it is the search string that a user decides to employ as a manifest representation of his or her needs[25].

A query is the search string that a user decides to employ as a manifest representation of his or her needs

This is not to say that it is necessarily a reliable representation of it; quite on the contrary. People are rarely able to define their needs with any degree of precision even to themselves, never mind communicating it to others (Morris, 1994). Consequently, searching becomes a highly imprecise enterprise. If we are going to provide relevant material to users, we need to find out what they are actually after, not just what they think they need. Consequently, queries are no longer there to provide the system with a search string, but rather to enable the system to decide what the user is looking for. This may sound as if it were the same thing, but a helpful search system does not actually look up instances of the query text as such – or even mentions of a query topic – but rather retrieves resources that seem to be relevant to the user's request, based on the evidence provided in the query, other user interaction, and contextual information. In other words, understanding what made users express themselves the way they did is the key to success.

Our goal first of all necessitates some further investigation into the search process as a psychological phenomenon experienced by the user (section 2.1). Were it merely a matter of issuing a query and receiving a response, search would be easy and simple both to perform and to understand, but this is not the case. Users typically approach a LAT because there is something they do not know, which necessarily means that they generally do not fully understand their problem. Consequently, they need to go through a complex psychological process of interaction with the tool during which their needs become progressively clearer to them and they formulate more adequate queries until, hopefully, they achieve a satisfactory result (Kuhlthau, 1993). Understanding how users construct and reconstruct their needs while the information received reconfigures their knowledge state and generates frustration and satisfaction is crucial; there is much psychology and construction of meaning involved, and this has repercussions for the way search tools should be constructed.

[25] However, there are many ways of representing a need, something that is the topic of Part III.

Search is typically associated with some form of information need (or other kinds of needs). User needs are often vague and always difficult to define. Nevertheless, users are themselves forced to come up with some form of description of their need in the form of a query – or at least something that will help satisfy it. How far the query is actually a reduced or distorted version of the need is not clear. The relation between need and expression is explored in section 2.2.

It is not immediately clear what needs are in the first place, what their relationship to the user's situation and task is, and what facet of a resource users are interested in. Section 2.3 presents the Faceted Stimulus model which was developed in order to allow discussion of such points.

Finally, section 2.4 is devoted to that other hard-to-define notion: *context*. This concept is rarely properly defined, and few people bother to spell out what it involves, as this is clearly an immense job. We shall be developing an understanding of context that seems to work for IR purposes at least, in order that we may jump into the dangerous seas of query log interpretation in Part III in the safe knowledge that our foundations are solid.

2.1 Undergoing the search process

Having defined the kind of tools with which this study is concerned and perhaps even agreed on the purpose of a search engine, it is time to look at the search process itself and explore what actually happens when a user searches for information. This section sketches the search process as experienced by the user. User behaviour is affected at least by the following cognitive, psychological and social factors.

- What the user thinks he or she needs
- The concepts possessed by the user
- The knowledge possessed by the user
- The linguistic/communicative abilities of the user
- What the user thinks can be achieved with the tool and resource collection in question
- What the user thinks is the right way to communicate a need to the tool
- What the surrounding society seems to think is the right way to go about searching

Behaviour is only secondarily affected by the actual workings of the system and the contents of the collection; this only happens through exposure to the system in a trial-and-error process where the user realizes what seems to work best. The same is true of the user's ability to use the system – a lack of knowledge of how the system works affects the way the user thinks a search is best performed, and this in turn affects the actual behaviour. In the end, it is the user's beliefs and abilities that govern behaviour. The concepts, knowledge and abilities of the user are hard to access and include in the interpretation process, but we shall have to deal with the others: User needs, meta-awareness and standard ways of communicating with a LAT.

2.1.1 The art of matchmaking

It is an important finding of prior research that search is not just a simple exchange but a complex psychological process (Kuhlthau, 1993). Searching for something affects the user's perception of the task and leads to restatement and modification of queries. However, the psychological process of searching in fact includes several phenomena apart from this emotional and intellectual struggle, among them the following.

- ❖ Relevance and applicability issues
    - Users' ability to assess what (ideal) resources would be needed, useful or relevant in a given situation
    - Users' ability to assess what actual available resources might be useful or relevant to the task at hand
- ❖ Knowledge and understanding issues
    - Users' initial (lack of) understanding of the relevant topic(s)
    - Users' ability to expand their understanding by incorporating new information
- ❖ Experience and skill
    - Users' expertise in using the tool
    - Users' ability to judge what the best search strategy would be in the circumstances

Traditionally, the basic communicative intentions of search engine users were taken to be the same across queries – something along the lines of *GET ME SOME INFORMATION ON* X – while the query was merely thought to refer to a topic of interest *X*[26]. The query topic would then act as a kind of stand-in for all documents on that topic. Here is a recent example of this line of thought, which demonstrates that it has remained influential to this day.

> Une requête est une formulation du besoin d'information d'un utilisateur. Elle peut être vue comme étant une description sommaire des documents ciblés par la recherche.
>
> 'A query is a formulation of a user's information need. It may be considered a summary description of the documents targeted by the search.' (Daoud M. , 2009, p. 13; my translation)

Thus, a document that was entirely "on the topic in question" would be the perfect match. The task of a search engine was accordingly to identify the resources in the collection searched that best matched the topic. It was soon realized that people would often not be able to provide the perfect query that would lead to a successful search result, so that the system might potentially have to help the user with **reformulation**; queries could be manipulated in ways that would lead to better results that would hopefully prove more helpful to the user in the end (Chang, Ounis, & Kim, 2006). Many queries were furthermore found to be ambiguous. Seemingly identical queries might represent different intentions, or at least different topics (Sieg, Mobasher, & Burke, 2007). By now we know that user intentions do actually vary significantly between searches, even when the same query expression is used and the same collection of resources is searched, quite apart from semantic ambiguity (Teevan, Dumais, & Liebling, 2008); sometimes people search for information on a topic, and sometimes they require answers to specific questions; some need help with practical problems (which can be very hard to convert into a query), and some simply wish to be taken to a certain homepage and use the query string as a way of locating it (Broder, 2002; Teevan, Alvarado, Ackerman, & Karger, 2004; Bergenholtz & Gouws, 2010). People have a wide variety of reasons for using search tools and as many interpretations of the queries submitted. The question naturally arises how we can possibly know what the user actually intended. Automatically recognizing intentions may well strike the reader as an impossible task, but fortunately we do have some data on the user's intention; data that are not currently exploited in full.

[26] Of course it may well be, as claimed by Hjørland (2010), that such a view has never really existed, in the sense that people really thought of this as being an ideal. However, early solutions do treat queries in this way, if nothing else then because there was simply no technological possibility of distinguishing between kinds of queries or constructing user profiles.

Today, matching is no longer believed merely to be a question of topicality (Harter, 1992). Nevertheless, the idea that the query represents desired document topics seems to have stuck subconsciously in people's minds (cf. the above quotation). Rather than accepting that not all users (indeed, as we shall see with one of the logs analysed in this thesis, in some domains vanishingly few users) search for topics in the traditional sense, researchers have largely focused on identifying different types and shapes of topicality depending on context, either by semantic disambiguation or query classification (see chapter 7)[27]. This has made discussions of the centrality of topicality go on for an unnaturally long time (they continue even today), in spite of much evidence to the contrary, because if a query represents a document topic then how can matching possibly not be about topicality? The answer must be that queries do *not* represent document topics, but user interests, or "interest topics", as it were, and that is not the same thing. In order to explain what this means, I shall introduce some examples from the Workindenmark job portal intended for foreigners looking for employment in Denmark. Some of the users of this system produced the following queries.

2.2	**ccna**	WID (89)
2.3	**no experience**	WID (21)

These queries were in all probability issued by people having earned a CCNA[28] certificate and looking for a job that requires no previous experience, respectively. They accordingly refer to aspects of themselves rather than properties of the job they are looking for. The documents sought are certainly not in themselves *about* CCNA and NO EXPERIENCE. Of course, in some sense the ideal documents do treat these topics, but they would at best be mentioned as incidental subtopics and would not be topics defined by the documents but rather imposed by the users and their interests. Thus, they actually represent user interests rather than document topics[29]. The topics are emergent features of the documents that are there as soon as someone looks for them, without being topics of the documents in

[27] In the interest of fairness it should be noted that search behaviour is likely to have changed significantly as new kinds of systems were developed and users experienced new possibilities. Thus, the amount of truly topical queries probably was greater in the past.
[28] Cisco Certified Network Associate.
[29] It is always possible to claim that a given query represents a mere wish to locate a document containing the string in question. However, as shall be argued in Part III, being satisfied with that explanation when others present themselves as viable merely hinders progress. It is simply not possible to help users who are only interested in string matching beyond the simplest of search engines, and so all IR research would be in vain unless further intentions are recognized. As also pointed out in Part III, there is no reason why it should not be assumed that a user may have more than one simultaneous intention, such as locating a document containing the string *ccna and* telling the system something of his or her own competencies with a view to finding a suitable job.

and by themselves. This is completely in line with constructivist ideas of information as something that arises in the moment of interpretation (Morris, 1994; Dervin B. , 2003).

> (…) it may be more useful to conceptualize human beings as information designers rather than information seekers and finders. (Dervin B. , 2003, p. 329)

However, sometimes topics seem indeed to be present more objectively (Behrens, 2005), and some users *are* interested in information on a major topic that is present without the help of the user as it were, but the point is that this is in no way necessarily the case. It follows that a query has the potential of telling us about the user's interests and intentions rather than merely about the documents sought, which opens a door to helping users find what interests them rather than simply what they expect to find. It is this door that we shall be peering through and opening up further in the course of this thesis.

2.1.2 Meanings and intentions

If we are to understand what the user is looking for or would benefit most from receiving, finding a way of decoding the intentions behind a query is paramount. Sometimes, the clue for such classification is in the query itself, which may contain expressions that strongly suggest a certain intentional interpretation. It thus pays to consider the actual expression more closely than simply as a way of establishing the topic. To some extent, such classification of query types has been carried out ever since Andrei Broder's (2002) paper on query types. However, despite the fact that the first search systems were developed by librarians, most researchers currently contributing to the literature on information retrieval have a background in computer science and consequently seem to treat the query as little more than a string of characters (Daoud M. , Tamine-Lechani, Duy, & Boughanem, 2009). Apart from the lexical meaning of query terms, researchers typically do not go beyond considering such superficial features as the number of terms in a query, the use of specific word classes (e.g. interrogative words, verbs, or proper nouns) and of certain lexical items that signal specific kinds of intentions (e.g. *download*, *buy*, *software*, and so on, which signal a so-called *transactional* intention). This is certainly a huge leap from *not* attending to such details, but it is very far from being enough. By taking a linguistic approach and applying a linguistically informed analysis to the user's linguistic choices it is possible to get much closer to users' actual intentions and hence to provide them with more helpful search results. The main reason that it has not caught on is likely to be the overwhelming focus on web search, where statistical methods seem in general to outperform detailed linguistic analysis, though even in this area there appears to be a reawakened interest in natural language processing now that the statistics are in place. Another reason may be the risk of falling into the trap of believing that understanding the intention behind a linguistic query is to understand its *meaning*. As pointed out by van Rijsbergen (1990), in search what matters is not really the meanings of sentences in queries and documents but intentions (on the part of users and authors) and usefulness (as regards resources), and it is important to keep these safely apart.

> In IR we do not seek an answer to the meaning of language, instead we seek a model that will enable the user to find information, that is, discover something she did not already know. Simply retrieving meanings is not enough; these meanings must carry information, hence this form of retrieval is inherently uncertain. (van Rijsbergen, 1990, p. 35)

Otherwise, we would expect ex. 2.2 to be a request for information on CCNA, which it clearly is not. We need to go beyond semantics in order to arrive at intentions, but semantics is an important step on the way to getting there, because it is only by knowing what *ccna* means that we can start to contemplate what the intention of issuing such a query might be in the relevant context. However, automating the inferential process from understanding the meaning of a string to a belief of recognizing the intention behind the utterance is a task fraught with difficulties.

2.1.3 Overcoming uncertainty

As pointed out by the philosopher John Dewey (1933), a person confronted with a problem in general only knows exactly what the problem is the moment he or she finds a way of solving it.

> If we knew just what the difficulty was and where it lay, the job of reflection would be much easier than it is. As the saying truly goes, a question well put is half answered. In fact, we know what the problem *exactly* is simultaneously with finding a way out and getting it resolved. Problem and solution stand out *completely* at the same time. Up to that point, our grasp of the problem has been more or less vague and tentative. (Dewey, 1933, p. 108; emphasis original)

It is thus no wonder that users of a search engine find it difficult to state what they need, the reason being that people typically consult a search engine because of experiencing a lack of information, i.e. there is something that they do not know or do not entirely understand. When a search session is initiated, the topic is therefore generally not fully understood and/or the user's situation and needs are not clearly represented in his or her own mind. A series of stages have to be gone through in which more information is obtained and the problem continually clarified. In this process, the query is often reformulated on the basis of the developing insights, and new searches are performed until a satisfactory state is reached, or the endeavour is given up due to a lack of useful results.

Based on the pioneering work of Dewey, and of clinical psychologist George Kelly (1963) and psychologist Jerome Bruner (1973; 1986), library researcher Carol Kuhlthau (1993) developed a six-stage model of the information search process. Each stage is characterized by a certain feeling in the struggling user (fig. 2.1).

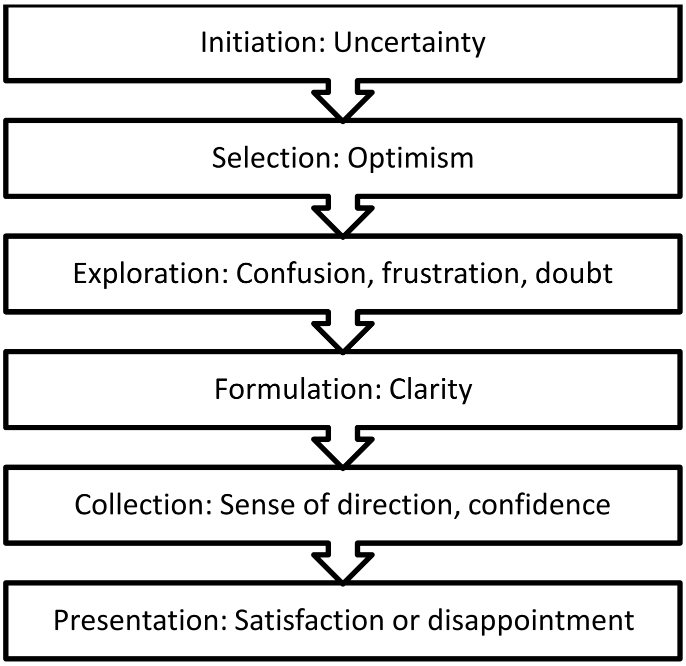


Fig. 2.1. Kuhlthau's search process stages.

The *initiation* stage involves the user recognizing a general need for information but not what might constitute the necessary information. The second stage is *selection*, in which the user identifies the general area of interest which might supply a solution if a search is made. This is followed by *exploration*, which is "often the most difficult stage for users and the most misunderstood by [library] intermediaries". The user does not know exactly what the problem is and hence is forced to look for evidence among sources of more general information on the topic. Through this, the user's understanding of the topic is improved, but this actually increases the user's confusion and lack of clarity about the problem, because only a small amount of the retrieved information contributes to the solution of the problem, the rest simply confusing matters as regards the task at hand and disrupting what the user thought he or she knew.

In search terms, the recognition of a **real need**, which may be at variance with the **presumed need** that instigated the search, is not always an easy process for the user to undergo; the user is forced to search rather blindly, and the inability to express precisely what is needed is often frustrating. The turning point is provided by the *formulation*, at which stage the user begins to acquire a focus gained through an improved understanding of the topic. Frustration ceases or decreases and a new optimism appears as search for a specific piece of information is initiated. The *collection* stage completes the now focused search, while *presentation* involves the final application of the retrieved information to the problem.

Kuhlthau's model describes certain kinds of search only; other types of search are less governed by uncertainty. Nonetheless, it does explain a lot of what is going on in a search process – especially in the more problematic types of search – and this psychological model of the search process is consequently at the basis of the Stage-based Intention model developed in Part III, which builds on the understanding acquired and the data gathered by these and other researchers through the ensuing decades. As our focus shall be on the construction of a query, the model deviates from Kuhlthau's findings in certain important respects, but the main idea is similar.

Many studies have shown that users behave in highly dissimilar ways, and certainly not usually in the way system designers – being naturally experts on using the system themselves – would expect (Navarro-Prieto, Scaife, & Rogers, 1999; Vakkari & Hakala, 2000). The most important lesson to be learnt from such studies is probably that a system should be able to cater for a wide variety of users and should allow vague and precise searches alike but treat them differently because the psychological need behind such queries is very dissimilar.

2.2 User needs and query construction

Our goal requires us to identify the **intentions** that lead users to issue certain queries. This is not a new undertaking; as we shall see in Part III, such studies have been going on for at least a decade. However, what exactly is meant by *intention* is rarely specified.

> [T]here has been little discussion of what is actually meant by user intent or what the theoretical underpinnings of the concept are. (Jansen, Booth, & Spink, 2008)

> [I]t remains a challenge to have a scientific definition of "intent" (...). (Manshadi & Li, 2009, p. 861)

Presumably, the intention behind a search derives from some sort of **need**, so we shall start by exploring how needs develop into substantiated queries. It is widely acknowledged that information needs change during the search process; users typically go through a series of stages in which the problem is identified, possibly formulated, and then turned into a query. The LAT responds by providing access to resources which would seem to be relevant to the user. Finally, users evaluate these resources (or the references to them shown in the results list) and determine whether they are really (likely to be) relevant to them in the situation at hand. This may prompt them to rephrase the query or even reinterpret the initial problem and so perform a new search.

The process of deciding on a certain query when finding oneself in a troublesome situation or facing a task has been modelled in a multitude of different ways. The end product of this process of query construction is a string that reflects a *presumed* or *compromised need*. The starting point is widely believed to be the user's *real need*, which is perceived by the user (often mistakenly) and then "reduced" further into a query. However, the user can hardly know in advance what his or her real need is. Were this the case, the task would have been much easier, but in fact people mostly only recognize their real needs when they receive the responses and evaluate them (cf. the quotation from Dewey above). This section looks at the various ways in which the development of a need has been modelled by various authors.

The traditional concept of *needs* pertains mainly to information tools, though it could be adapted to other kinds of tool by replacing *information needs* with more general **response needs**. However, we shall concentrate on information needs at present for simplicity. In general, only researchers who deal specifically with the nature of the concept deign to define it, and when others do, the definitions tend to be behavioural, for instance defining user needs as "the perceived need for information that leads to someone using an information retrieval system in the first place" (Schneiderman, Byrd, & Croft, 1998). This necessarily restricts user needs to the perceived needs at the beginning of the search, which precludes any sense of a real need or even a dynamically evolving one. This is also true of some definitions which are more cognitively adequate, at least in the sense that they do not allow for the recognition of real needs not perceived as such by users themselves.

> The approach I have taken to describe an 'information need' is an awareness of a state of 'not knowing' – or some conceptual incongruity in which the learner's cognitive structure is not adequate to the task. (Ford, 1980)

This appears rather to be a definition of the state that people are in when a need arises than of the need itself. As explained in section 1.1.3 above, we shall adopt a more technical definition of needs which connects with the model of relevance suggested in Part II. Since the relevance of a resource is subjective and dynamically evolving, this definition covers all stages of the developing need. The definition is repeated here for convenience.

> A <u>need</u> is a relation $n(D,C)$ between a context C and the resources D that are relevant to the user $u(C)$ in the context C. The relation is such that D comprises exactly those resources needed by the user in context C.

Users are here considered a function of their context, since a sufficiently detailed description of the context will include the user. A need is accordingly not the relation between a user and a document set, but between the context in its entirety – including the user – and the ideal response document set. However, there can be no context without a center to which it is contextual, and thus C necessarily presupposes a perspective from which the situation is experienced. That is why such a personal and individual thing as a need can be ascribed to context rather than to an actual user in particular.

Saracevic *et al.* (1988) claimed that the concept of *information need* had been superseded by a new and better understanding that involved studying the problem-generating situation.

> For a long time the predominant concept around which models revolved was the concept of information need (...). Slowly, modeling changed to that of problem orientation, viewing the problem behind the question, rather than information need as central to information seeking context. (Saracevic, Kantor, Chamis, & Trivison, 1988, p. 163)

It is true that the 1980s saw a lot of work on user- and situation-oriented studies, but to this day *information need* remains an important concept in IR research. The definition of *need* that we have given above explains why this is so; information needs consist of a relation between a context (including the problem-generating situation) and a set of desired or needed resources, the latter being closely related to the traditional notion of information needs. Both halves of the concept are equally important and correspond to what we shall be calling **Stage 1** and **Stage 2** of the development and clarification of needs. We shall see in Part III that users may refer to both stages with their queries.

2.2.1 Query construction models

It is important to understand that the task and the situation in which users find themselves when confronted by it are really one.

> [T]here is not at first a situation *and* a problem, much less just a problem and no situation. There is a troubled, perplexed, trying situation, where the difficulty is, as it

> were, spread throughout the entire situation, infecting it as a whole. (Dewey, 1933, p. 108)

The problematic situation is one characterized by "a certain incompleteness in [the user's] picture of the world – an inadequacy in what we may call his 'state of readiness' to interact purposefully with the world around him in terms of a particular area of interest" (Mackay, 1960). This situation is perceived and interpreted by the user in order to identify potential information or actions which might solve the problem. If relevant information is found, this contributes to a more profound understanding of the problem, which the user may then be able to tackle, often by acquiring more precise information. Users' needs thus stem from the situation in which they found themselves as the need arose, though the precise relation between problematic situations and needs is likely to be highly varied. The need is often thought to derive directly from the task by being a kind of reduction of it in which everything that is not strictly relevant to the solution of the problem has been trimmed away.

> The user has a problem, perceives it, interacts with an IRS (and maybe a human intermediary), expresses his information need in a request, formalizes it into a query, examines the retrieved documents, reformulates the query, re-expresses his information need, perceives the problem in a different way, and so on. (Mizzaro, 1997)

This may or may not be true, but at any rate the common conclusion (the strong version) that the query is in the end some kind of reduction of the representation of a task situation in this way does not seem to hold; either the query is not a true representation of the user need (not even an imprecise one), or the need is not trivially derived from the task, or both. This question of course also depends on the definition of the task, i.e. whether it is considered objectively as seen by an affectively neutral outsider or as experienced by the troubled user. As pointed out by Ingwersen and Järvelin (2005a, p. 72), "how the need originally arises and how this relates to the person's situation or tasks remains unanalysed". We shall take a look at four different models of query construction from an initial sense of need.

● Taylor 1968

Our first example comes from the tradition of librarianship. The model is taken from Robert S. Taylor (1968) and reflects the reader-librarian dialogue. Taylor reckons with four stages (fig. 2.2).

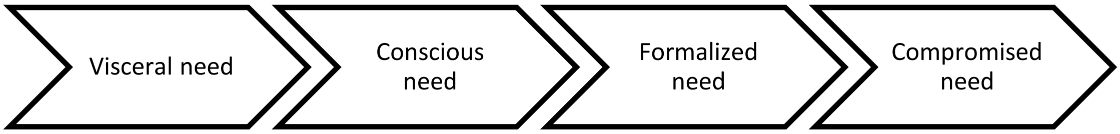


Fig. 2.2. The query construction process according to Taylor.

The model is psychological and behavioural in that it does not portray the representation of a need as a reduction from an actual disembodied need to a substantiated search string, but rather as a narrowing down from a confused and partly subconscious state of mind to a stage where the inquirer is able to verbalize the need. This is in stark contrast to most non-library models.

The **visceral need** is the actual but unexpressed need for information that made the user consult a librarian. It may not be a conscious need, but simply "a vague sort of dissatisfaction", and it is frequently not directly expressible in linguistic terms (Taylor R. S., 1968, p. 8). Note that this is not what in IR research has come to be known as the **real need**; the visceral need is psychological, and the satisfaction of it may well turn out not to solve the problem. The **conscious need** is a mental image or description of "an ill-defined area of indecision". At this stage, the user typically needs to talk to someone else in order to clarify what the crucial points are, so as to be able to express the need, which is where the librarian enters the stage. The **formalized need** is a rationally expressed question, i.e. a request for some definite information. Interestingly, at this stage the dialogue with the librarian changes. The librarian is now considered to represent the information system with its formal classifications, rather than a colleague with an understanding of the topic in question. The last stage in Taylor's model is the **compromised need**, where the user anticipates the categories as well as the shortcomings of the system and trims down the description of the need to one that there is some hope of satisfying. Often this compromise takes the form of the *label effect* (Ingwersen, 1982; Ingwersen, 1992; Ingwersen & Järvelin, 2005a, p. 296ff); people tend to substitute full descriptions of what they need with simple labels that refer to only one of the involved concepts, thereby overly simplifying the problem statement and narrowing their chances of pinpointing something truly relevant, and even affecting their own understanding of the problem in a negative way by *anchoring* all succeeding reformulations to the labels first chosen (Tversky & Kahneman, 1974; Blair, 2002b). The *compromised need* is of unquestionable importance to IR research, but it tends unfortunately to be forgotten.

Cooper 1971

The second model is a traditional approach to non-library oriented IR research. It owes its character to the undertaking of constructing inferential systems. Unlike search engines, such systems basically answer questions, and these are likely to have been expressed in linguistic form before being formalized and submitted to the system. Consequently, the model owes a lot to the library tradition. The example (fig. 2.3) is taken from William Cooper (1971).

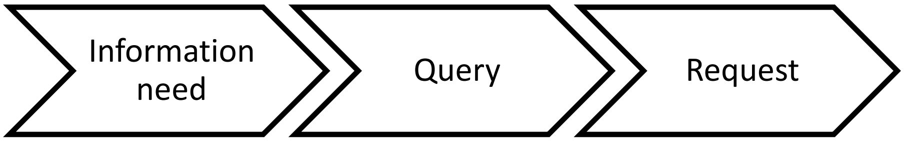


Fig. 2.3. The query construction process according to Cooper.

The **information need** in the strict sense intended by Cooper is a psychological state corresponding to Taylor's *visceral* and *conscious* needs. Unlike Taylor's model, however, this one is viewed from the point of view of the system, and thus Cooper does not care whether the need has formed into a conscious mental image of the problem or not; the information need is a general pre-expression stage. In Cooper's terms, the *query* is an attempted formulation of the need as a question, which may or may not cover the user's real needs. It thus corresponds to Taylor's *formalized need*. What Cooper calls the *request* is the actual string submitted to the system, which rather confusingly has later come to be known as the *query*. This is in fact a *compromised need* (cf. Taylor); it has been adapted to the anticipated limits of the system. The difference between the *request* and Taylor's *compromised need* is that the latter is still a psychological state, while the former is "the input actually submitted to the retrieval system and upon which the system bases its search" (Cooper, 1971, p. 21). The transformation from query to request in Taylor's terms can be performed by the user, a human intermediary (Taylor is, after all, from the library tradition), or the system. Ingwersen and Järvelin (2005a) adopt this model, except that they follow current practice in swapping the terms *query* and *request*. Thus, to them the request is "the formulation of the information need or the underlying states of intentionality, as perceived, and provided at a given point in time by the actual searcher to an IR system or other information sources", while the query is "a transformation of a request formulation made by an actor himself/herself, an intermediary or an interface in order to interrogate an IR system's information space, in concordance with the system's indexing and retrieval algorithms"

(Ingwersen & Järvelin, 2005a, p. 20f). Clearly, this tradition lives on in library research, for good reasons.

● Saracevic 1996

The third model is the one inferable from Saracevic's (1996; 1997b) *stratified model* of IR interaction. Just as the information content of a resource is constrained by the software in which it is couched, which is in turn constrained by hardware capacities, the needs generated by a situational task are constrained by the user's affective intent, which is again constrained by his or her cognitive knowledge structure and capacities, which in turn is constrained by the characteristics of queries (Xu & Chen, 2006). The idea that the need is *constrained* by different factors explains in an elegant way why there should be such a thing as different stages of need development. Saracevic's model is highly psychological, but as a cognitive explanation it is rather naïve. It sets the stage for a cognitive interpretation of need development, but does not really provide it; instead one is referred to Ingwersen's (1992) work. The model is neither reductionist nor behavioural, in that the constraints apply non-sequentially because they are all present from the start. Consequently, it does not fit into the graphic model of a developing need, which is just as well, because the schematic illustrations are probably much too simplistic to be at all like the real process. Section 2.3 looks closer at the *Stratified model* and introduces a new *Faceted Stimulus model* which is based on similar ideas but explains user intentions and reactions better than the former.

● Mizzaro 1998

The fourth model is typical of recent IR research, apart from those who embrace Saracevic's Stratified model. The example is based on Mizzaro (1998) and is a further development of Cooper's model. It, too, inherits the linguistically expressed version of the need from the library tradition. However, it follows current practice in swapping the terms *request* and *query* so that the request is the supposed linguistic expression preceding the actual query submitted to the system. Furthermore, the fact that users are unable to clearly define their own needs is acknowledged by recognizing a **perceived need**, which is a further reduction of the **real need** derived from the task (fig. 2.4).

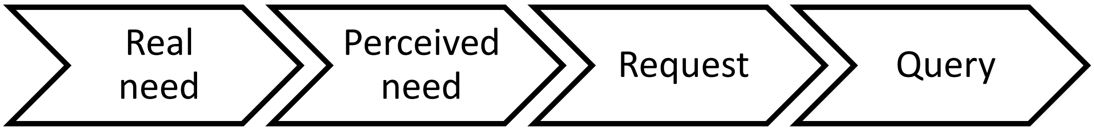


Fig. 2.4. The query construction process according to Mizzaro.

Mizzaro regards the query (i.e. the actual string) as a representation of the request or information need, the request as a representation of the information need expressed in a human language, and the (perceived) information need as a representation of the problem in the mind of the user (Mizzaro, 1997). In other words, the whole chain is a process of reduction, from the actual situation to a highly stylized representation of it (with reference to the user). In this way, it becomes not so much a psychological model of a process from uncertainty to assertion as a question of decreasing reliability with regard to the real (true) need as the need representation is repeatedly reduced and possibly misinterpreted. Still, Mizzaro's is a psychological-behavioural model in that it incorporates a verbal request which never reaches the system, yet it also includes a real need, which is non-mental, and the *query* is clearly intended to be a system-interpreted string rather than an expression produced by the user. It is thus of a somewhat mixed character.

Mizzaro's model is perhaps rather more suitable when it comes to reconstructing the real need from the query (i.e., with the arrows pointing in the opposite direction) than as an explanatory framework as regards the cognitive process from real need to query. If the query can be expanded into a full linguistic request which can be interpreted as representing some perceived need, then perhaps the real need can be inferred from this information. Turning the model around and trying to explain the psychological process in the same way does not do, however. For one thing, the perceived need may be distinctly at variance with the real need and not just a reduction of it, and such erroneous need perception is not accounted for by Mizzaro's model. It also seems exceedingly unlikely that the query should in general be a reduction of a full linguistic expression, not least because users have such a hard time explaining what they need when asked to do so. Rather, the query should be considered to be that linguistic expression itself, and the syntactically impoverished format in which it is mostly found (whole sentences are rarely employed) is a consequence of the non-typicality of the addressee, the tool, which is not expected to be able to interpret complex descriptions, in addition to the inherent difficulties that users experience in expressing or even understanding their needs (cf. Saracevic's constraints). Rather than phrasing questions and reducing them to a query, I take it that LAT users phrase reduced questions or requests *in the form of queries*. The model advocated here is thus more closely related to the one suggested by Saracevic.

2.2.2 Real needs

The so-called **real need** is a hypothetical one of equally hypothetical information which might potentially lead to the solution of a task. By definition, the satisfaction of the real need solves the user's problem. The real need is thus not so much a product of the task as one construed from the available information in the interpretation phase[30]. Since users in general have no way of knowing what information is potentially available and hence what might possibly contribute to the solution of the problem, they cannot know what their real needs are. Indeed, the problem solved by the real need may not even be the one that the user set out to tackle. If the user ends up satisfied, his or her real need must be considered to have been satisfied. In conclusion then, the perceived need cannot be a reduction of the real need, but rather is entirely independent from it, the only potential connecting factor being the actual task. Consequently, real needs are not part of query construction at all, but rather belong with relevance evaluation after a response has been obtained. Only upon interpreting the results after the search has been completed does it become apparent what information was actually helpful (Barry, 1994; Ingwersen & Järvelin, 2005a, p. 72). A need that is not in principle reconstructible from the combination of query (and other information provided by the user) and information in the document collection is not relevant to the system since it cannot be satisfied. The concept of *real need* is replaced here by a situationally induced type of relevance to be discussed in more detail in Part II.

The reader might be wondering at this point what the start of the process could possibly be if not the real need. The answer that we shall put forth is that users who are in a problematic situation initially find themselves in the grip of a *visceral need*, which is a psychological awareness of an indefinable lack of information. This may with time mature into a presumed need, which reflects the way the user perceives the task at hand. In simple cases this will happen fast, in others it may require hours of study. In the end, the real need may turn out to be something entirely different, and all other kinds of need representations are mere proxies for it.

2.2.3 Presumed and compromised needs

In the received models, the query is derived more or less directly from a perceived need. However, the presumption of a need stems from the experience of a task, and the needs arising in such situations are not necessarily informational, which means that they often cannot be satisfied by a LAT. Users know

[30] The notion of *availability* is here meant to include users' *awareness* of availability (Barry, 1994).

very well that search engines and other information tools cannot be expected to solve just any problem; they appreciate that there are limits to what the tool can do, and they modify their perceived needs accordingly before approaching it. The unmodified perceived need we shall call the **presumed need**. For instance, if I have a task which requires me to go to Valencia for a conference, my presumed need might be *TRAVELLING TO VALENCIA*. I cannot expect a search engine to fulfil this need directly, so I settle for an intermediate goal of searching for tickets to Valencia. Even this may turn out to be less than straightforward, however, because *tickets* is a rather vague term, referring to different kinds of transportation as well as entertainment. In principle, there is no reason why I should not be able to obtain the needed information, but in practice it may be difficult to find unless I restrict my search to something more specific like *flights to Valencia*. This is a case of what we have called **meta-awareness** (see also section 10.2); the user must estimate what expression is likely to give the best results based on the expectations as regards the contents of the full document set and the abilities of the system. Paralleling the difference between the need of actually getting there and simply buying the tickets is the one between the *presumed* and the *compromised need*. The latter is taken as a basis for the formulation of a query that should enable the tool to return the anticipated result. It is very much akin to what Taylor (1968) understood by the term, which has accordingly been adopted here. The compromise consists in a combination of recognizing the limits on what the tool can deliver and meta-awareness reflections on what behaviour is likely to result in the best outcome when interacting with a particular tool (fig. 2.5).

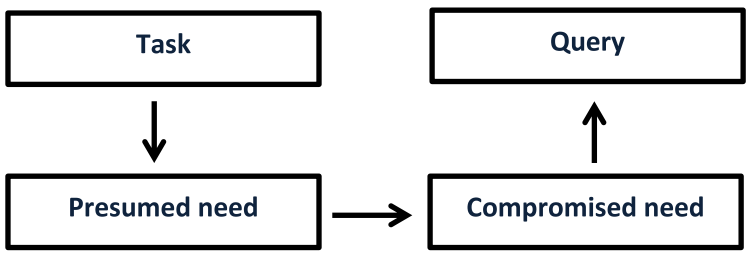


Fig. 2.5. Needs from task to query. The presumed need is derived from the task by subjective evaluation, and is reduced or moulded into a compromised need suited to the requirements of the LAT. Finally, it is formalized and substantiated as a query.

Since the processing of presumed needs into compromised needs is indeed a kind of reduction or distortion, we can only get to the presumed need by understanding the task as experienced by the user. The compromised need, on the other hand, is reflected in the query, assuming that the user has successfully formalized it, which is of course in no way certain.

The *real need* is realized by the user after the search has been made (according to relevance criteria discussed at length in Part II), in the event that some truly relevant information has been found which solved the problem. In that case, the real need is constructed in the user’s mind on the basis of the search results and the task. In fact, even the user's conception of the task may change in the process (Kuhlthau, 1993). The construction of a real need may leave the task in such an altered light that the real task may turn out to have been something completely different from what it was initially thought to be. Hence, the presumed need is derived from a **presumed task**, while the **real task** is inferred from the real need (fig. 2.6).

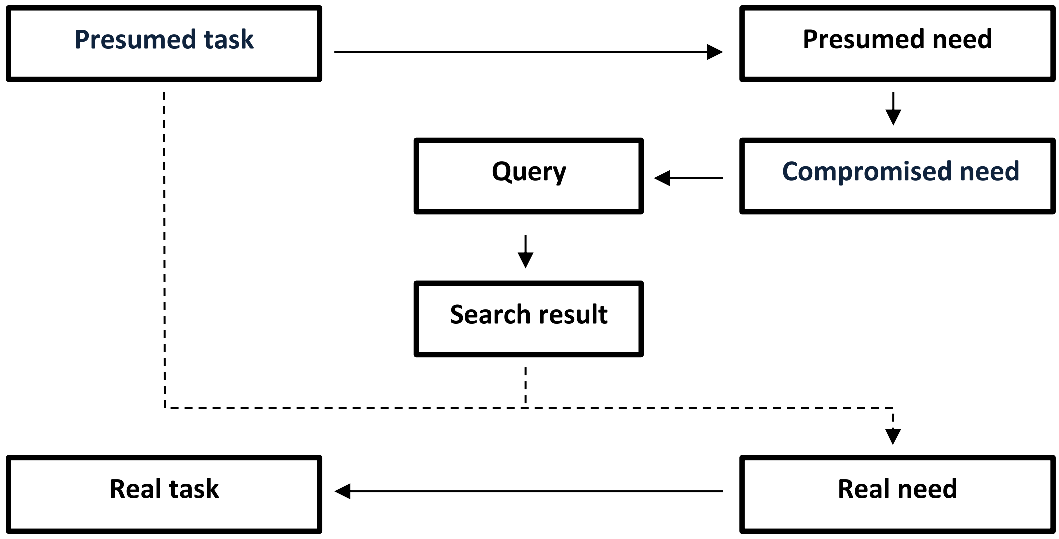


Fig. 2.6. Recognition of needs and task as a constructive process during search. The recognition of a real task differing from the one initially presumed to be the issue frequently leads to a new query in which the task has been reconceptualized in light of the previous results.

In this light, there are two main possibilities as regards improving search tools. One is to be content to reconstruct the compromised need in greater detail and attempt to satisfy it, and the other to make a stab at reconstructing the presumed task (and through it, potentially both the presumed and the real need). The latter endeavour necessitates the addition of contextual information from an external source[31]. Either way, it is requisite that the intended interpretation of the query is understood to any extent possible, and even that is clearly problematic.

2.3 The faceted message

At this point it would probably be a good idea to look closer at the resources that users search for; after all, they must presumably be a vital part of the need. The term **document** is used here, as is often the case, of any kind of information **resource**, whether in the form of text, sound or video (Schamber,

[31] In accordance with this, Ingwersen (1996) suggests the addition of information on the problem space and the work task or task domain as a reasonable solution.

1996; Buckland, 1997). Furthermore, the terms *document* and *resource* are in general used interchangeably to mean any form of resource returned by the LAT, and to which it provides the user with direct access. A resource may thus be anything that the user receives in response to his or her query. Tradition makes it natural to speak mostly of documents, even though there are such other kinds of responses. Already, it has become common to treat images and videos as documents on a par with written media, and documents no longer need to be physically delimited in time and space but can be dynamic (Schamber, 1996). The step from there to a fully general concept of documents as being simply responses (including actions) is a natural one, even though the term may feel somewhat contrived because of its heavy connotations of text, or at least digital files.

Some of the antics of user behaviour are illustrated by Tefko Saracevic's *Stratified model of information retrieval interaction* (1996; 1997b). The idea behind this model is that interaction happens on several levels simultaneously. Behaviour is observable on the *surface level* but at the same time is governed by interaction with information objects on the *cognitive level*. This idea was probably inspired by Peter Ingwersen's work (1992), which he acknowledges to be one of two outstanding developments in IR theory at the time, the other being Belkin *et al.* (1995). He also proposes an *affective level* to account for intentionality and other affective developments in the search process. As emotion is reaction to an input, this level can be considered to involve interaction with the search process as such. Finally, he suggests a *situational level* at which users interact with the task situation, develop a need and construct a query.

In the following, a more detailed way of considering resources as consisting of several **facets** is introduced. It is felt that this is necessary since user intentions may vary in their relation to resources. Both resource documents and queries are complex entities consisting of strings containing linguistic messages, and hence an IR system needs to be able to interpret several facets of them.

In spite of Saracevic's intense interest in the phenomenon of relevance, the Stratified model is in fact not a model of relevance, but of *information retrieval interaction*. It focuses on and attempts to explain user experience and behaviour rather than resource informativity and usefulness. The intention seems to have been to construct a model that would also explain relevance, but in this he does not really succeed (see discussion in chapter 4). The **Faceted Stimulus** model introduced here deviates from Saracevic's model by focusing instead on the information that is at the centre of the search process and only secondarily explaining user experience and behaviour as side-effects of the hunt for relevant information. This gives rise to a model that is also able to explain relevance assessment with only a few additions to explain the interactions between the "strata". The model is somewhat related to Saracevic's idea of strata or levels, but we shall choose to call them **facets** or **dimensions**. The term *facet* is used of the various layers of resources, users, authors and situations; *dimension* is used later on when we wish to talk of the phenomenon as one that cuts through the notions of resource, context, relevance assessment and query formation and makes all of these phenomena interact as one big machinery.

2.3.1 Resource and query facets

A document is firstly a (usually digital) object which may or may not contain linguistic substance[32]. If it does, it also contains linguistic descriptions of semantic worlds. Some users are most interested in the substance, the actual string, while others only care about the semantic significance of the text, or even the usefulness of the text in a particular situation. In fact, resources will be shown to have six different facets of which physical objects (the **substantial** resource[33]), strings (the **formal** resource) and the semantics represented by them (the **representational** resource) are the least abstract (fig. 2.7).

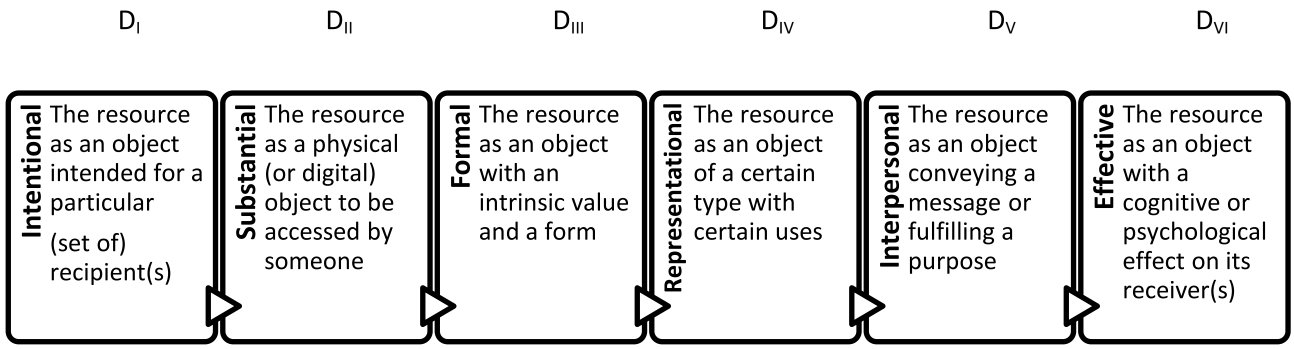


Fig. 2.7. The facets of a generalized resource. The terms *representational* and *interpersonal* have been borrowed from Functional Discourse Grammar (Hengeveld & Mackenzie, 2008). Their term *morphosyntactic* was dispreferred because of its heavy linguistic connotations and was hence replaced by *formal* in order to allow for non-linguistic resources as well. Note that it will be found in Part III that two additional facets of an abstract nature may well be called for – one between dimension II and III representing the type of resource and another between dimension IV and V representing the type of communicative or telic contribution it makes. This requires further research, however, and hence the numbering I-VI has been retained here, pending further investigation.

The others represent the document as a piece of communication (the **interpersonal** resource), as an object intended for a reader (the **intentional** resource), and as an object with a cognitive or psychological effect on the recipient of the resource (the **effective** resource). The full significance of the facets will only become clear in Part II of the thesis where each facet will be shown to be evaluated separately for relevance. In fact, the entire model of relevance suggested there centres on these same

[32] Latterly, search engines have been constructed that allow users to search for and in images and videos using linguistic queries (Peters, et al., 2006; Arampatzis, Zagoris, & Chatzichristofis, 2011). Thus, LATs are not restricted to linguistic resources, even if they are restricted to linguistic queries.
[33] Terms such as *the X resource* and *the X facet of a resource* may be used interchangeably.

six dimensions of resources. The suggested facets are highly abstract, as they need to be in order to encompass linguistic resources, images, sounds and videos, non-digital objects, and perhaps even actions. We shall briefly describe each below. The dimensions are numbered I-VI. It will be shown in Part III that there may be reason to erect two additional facets – one between dimension II and III representing the type of resource and another between dimension IV and V representing the type of communicative or telic contribution it makes. This requires further research, however, and hence the numbering I-VI has been retained here.

● **D_I – The intentional resource**

Most documents are created by someone with an intention for them to be read by a specific kind of user, though the degree of specificity may vary considerably. The intended recipient may be a specific set of people (or even an individual), a group of users characterized by some commonality, or any person in a certain kind of situation, or just anyone who happens upon the document.

> (…) one must study the origins of the various document types, how they are produced, for which users, and under what economic constraints. It is recognized that producers of documents generally have specific consumers in mind (…) (Brier, 2008, p. 45)

The intended target group may be crucial to its usefulness to a given reader, because a document intended to help the user in a particular situation is likely to be far more useful than one that merely happens to treat the topic in question. Also, if someone intended a specific user to read a document, that document will be relevant to the user in question by definition[34]. On the other hand, if the user is forbidden from accessing the resource, suggesting it is clearly not helpful. The intentional resource is a prelinguistic and even "pre-physical" concept; its nature is like that of a *gift* considered as an abstract entity irrespective of what is actually given.

● **D_{II} – The substantial resource**

The resource is of a certain type and format which allows it to be accessed by some recipients only. A resource may be inaccessible due to its format, its cost or geographical distance from the user. When users perform a pure string search in order to find or refind a particular sequence of characters, they target the substantial resource. This facet is non-linguistic, so if a

[34] Other circumstances may, however, render the act of suggesting the document to the user irrelevant; it may for instance be inaccessible, or written in a language not understood by the reader, in which case it is useless (unless such obstacles are done away with, for example by translating the document).

string occurs as a word in more than one language, they cannot be distinguished, and even if it just happens to occur in a non-linguistic stretch of characters, it is still the same string.

● D_{III} – The formal resource

Language enters the stage in the formal resource facet. Here, identical strings in two different languages clearly represent different words. On the other hand, inflectional variants and perhaps synonyms may be considered to represent the same word in this facet. With non-digital objects, certain formal characteristics that do not relate to the telic function of the object, such as its size and weight, relate to the formal resource facet. One might intuitively think that such features would belong with the substantial facet, but the substantial resource is about being able to receive and access the object, whereas the formal resource deals with the actual handling in the sense of making use of the resource. If the object is too big to be sent to the user, this is indeed a question of the substantial facet, but if it is merely too big to be suitable for the recipient (who might be a child, for instance), then it is a matter for the formal facet, because it interferes not with access but with use. A D_{III} text can only be read by someone who can understand the language in question, just as a D_{III} object can only be properly understood, valued and used by someone who knows what it is.

● D_{IV} – The representational resource

The representational resource contains the semantics of a document, and the typical characteristics of an object. Here, related terms may be considered to relate to the same topic, even though they are not synonymous but only ontologically co-occurring.

In a prototypical topical query, the user is interested in this facet of the resource[35]. With objects, the representational resource represents their type and the purposes that they are intended to fill, but only in a general sense with no reference to the situation that the user is currently in. The representational resource is not a text as such, but rather is the content created in the reader's mind when reading a linguistic document. The representational resource is thus alinguistic.

[35] This is not the defining feature of such searches, however. In fact, it is neither a necessary nor a sufficient characteristic of topical searches. As we shall see in Part III, what makes topical searches special is that they target a predefined topic constructed in the user's own mind and labelled in the query. Users looking for topical information may access this in other ways, such as by asking a question or entering an entire passage to be retrieved in a document. True topical searches, on the other hand, may well represent a wish to find, say, an interpersonal or effective resource rather than merely a representational one.

● D_V – The interpersonal resource

Some queries demand more than topical relevance. If you find yourself in a certain problematic situation, a text on the topic may not be enough; you may need a text that is particularly useful to you in that situation. In such cases, the interpersonal resource facet is called upon. It represents the resource as conveying a message or fulfilling a purpose to someone in a real-world situation. Note that for a resource to be particularly useful to people in a certain kind of situation must not be confused with the author intending it to be useful for people who find themselves in such a situation. The latter relates to the intentional resource and focuses on intention rather than utility. The situation is merely a characteristic feature defining the target user group of the intentional resource, while it is central to the interpersonal resource.

The nature of the interpersonal resource is similar to that of a *resource* that one may draw on considered as an abstract concept irrespective of what it actually consists of. While the representational resource covers general characteristics of the resource, including what it can be used for in general, the interpersonal resource has the characteristics that make it useful in a particular situation with an actual user involved. Thus, the former deals with function and the latter with use.

● D_{VI} – The effective resource

Finally, the aesthetic characteristics of a resource – such as layout and interestingness – pertain to the effective resource facet. Here, utility is set aside for a moment and the desirability of the resource is considered for its own sake. Sometimes, looks, style and the ability to engage do matter; then people may target the effective resource. In *serendipity search* especially, i.e. when people browse for entertainment, the effective resource is of key importance. The effective resource is a psychological entity constructed through, but not constituted by, (among other things) the aesthetic and manipulative features of the resource, including the contents of linguistic text. Just as the representational resource consists of the semantics rather than the signs coding them, so the effective resource consists of the effect created by the resource in the reader's mind rather than the actual features causing them. Even the cognitive effect of being informed by a resource is part of the effective facet, while the actual information of course pertains to the representational and interpersonal facets.

The six facets of a resource are bundled together into one complex object that users may search for. The question naturally arises whether users are in general more interested in certain facets than others. The answer appears to be that this varies with the search aim. With linguistic documents, people who wish to find (or refind) a site where they can perform some task may search for the substantial resource. Others are interested in the formal resource – they search for precise wordings in order to find (or refind) a stretch of text, such as a digitalized novel or a blog post. Some are interested in the representational resource, because what they are after is information. Many would be more helped by the interpersonal resource, needing information that is useful in a certain situation. Finally, some would benefit from locating an intentional resource specifically provided for people like them, while others simply want resources that they find pleasing in some way or other. All of these “customers” can be helped with different IR tools – or with one flexible tool that is able to adapt to the particular needs and wants of a given user. Here is one example.

2.4	**Copenhagen cafés**	*

If someone were to enter this query in a search box, they might be interested in getting to a specific page which they knew to carry this heading (the *substantial resource*), or they might be looking for any page with this wording (the *formal resource*) in the hope of finding something useful. Alternatively, they could be looking for any page containing information to this effect (the *representational resource*). This would be the more probable in this case, since not only documents with the exact wording *Copenhagen cafés* would be helpful to them. Finally, they might be looking for any page specifically designed to provide users such as them with a list of cafés in Copenhagen. In this case, they might not even need to see the actual page – in fact, it need not exist as a substantial document at all, as long as the tool were to provide them with a list of cafés in Copenhagen compiled by someone (the *interpersonal resource*). Such a virtual list might of course alternatively have been compiled automatically by extracting information from websites (a so-called *mashup*). Even in this case, there is an intention behind, as someone (the publisher, or **selector**[36]) instructed the system to compile the page dynamically with the intention of informing potential readers.

Queries are likewise a kind of dynamic documents, though usually very short ones. They accordingly have facets just like resources. We shall content ourselves with introducing four query facets, since intentional and effective facets are hardly useful in this connection, even if they are potentially meaningful. The **substantial query** consists in the actual characters submitted, whereas the **formal query** is an expression involving lexical items pertaining to some human language. The **representational query** contains ontological information relating the lexemes to each other and to

[36] The term *selector* originates with Ingwersen and Järvelin (2005a, p. 208).

other concepts. Finally, the **interpersonal query** incorporates the intention of the user to convey a message to the tool (fig. 2.8).

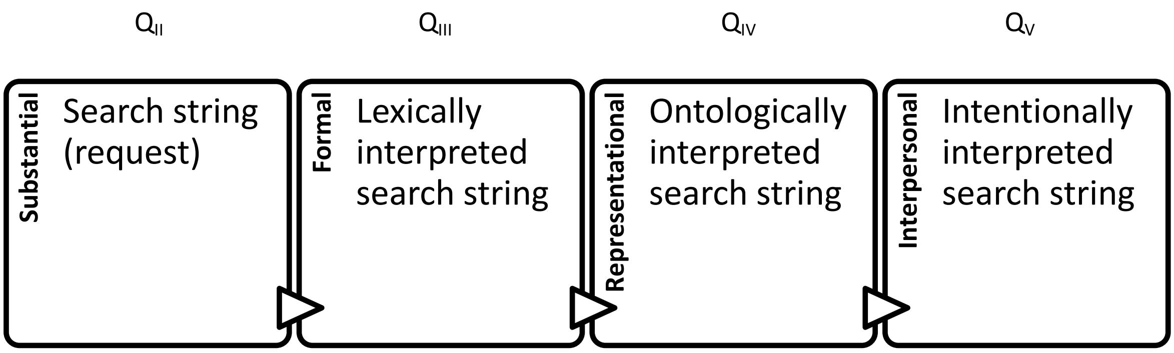


Fig. 2.8. The facets of a query.

In principle, you can only search for a document using a query that belongs to the same facet as the resource sought. If the user is looking for a certain string, for instance, ascribing semantic interpretations to the query as a linguistic message makes no sense, and certainly is not helpful. On the other hand, if the user is looking for advice in a given situation, searching with a string can at most be a proxy for an intentional query (i.e. a query enriched with an intentional facet). Queries are not just strings of characters but do convey intentions, at least to the extent that the system is able to interpret them. As soon as information retrieval tools are made so sophisticated as to be able to help users beyond performing mere string searches, it becomes important for the system to be able to tell the difference between facets of varying "richness" (i.e. the simplest query is of facet level II, and levels of higher orders represent enrichments of the query).

By way of an example, a query such as the following can be interpreted in many ways.

2.5	**sky**	*

In a Danish-language environment, the user might be supposed to be interested in texts on clouds (*sky*). In that case, the query would have at least a representational facet. If it didn't, documents containing the homonym *sky* 'shy' would be relevant too, because the search would be one of looking for the formal word *sky*, with no reference to the semantics. If there was no formal facet either, English-language documents on *sky* would be relevant as well, as would documents containing the letters *s*, *k* and *y* in succession by pure accident. Unfortunately, it is not possible to determine what facets a user intends to include without asking for explicit feedback, since having facets is an intrinsic feature of linguistic messages. Collecting feedback on the facet intended is probably not advisable, either, as users are likely to be thoroughly baffled by the question. A "dynamic IR" solution could be to support a diversified result presentation and return results relating to various facets when they are available.

2.3.2 Authors and interpreters

The informational content of a document is conveyed first by the **author** who wrote it, and then by the one who published it (which may or may not be the same person). Both may have an intention of conveying the message for some reason, and this intention may be the same for both parties, or it may differ. We shall sometimes refer to the author and the selector together as **informers**, in order to stress that they both convey the message in the document, and who actually wrote it is less important since we are more interested in the intentions of conveying a message to some addressee than the intention of producing a document.

User facets		Query facets	Resource facets		Author facets
U_I	➡	(Q_I Intentional query)	D_I Intentional resource	⬅	A_I
U_{II}	➡	Q_{II} Substantial query	D_{II} Substantial resource	⬅	A_{II}
U_{III}	➡	Q_{III} Formal query	D_{III} Formal resource	⬅	A_{III}
U_{IV}	➡	Q_{IV} Representational query	D_{IV} Representational resource	⬅	A_{IV}
U_V	➡	Q_V Interpersonal query	D_V Interpersonal resource	⬅	A_V
U_{VI}	➡	(Q_{VI} Effective query)	D_{VI} Effective resource	⬅	A_{VI}

Fig. 2.9. The various producing role facets of users, queries, resources and authors. The arrows denote production processes.

Authors as a rule have separate (but largely simultaneous) intentions of producing each of the facets of a resource. We shall distinguish these intensions as emanating from different facets of the author: The A_{VI} author affects the recipient by creating the effective D_{VI} document, the A_V author produces the interpersonal D_V document, the A_{IV} author constructs the semantic D_{IV} document, the A_{III} author formulates the formal D_{III} document, the A_{II} author creates the substantial resource, and the A_I author intends to create something for a certain set of recipients. Users, or in this case readers, are similarly faceted; the U_I user inhabits a social world in which resources may be intended for certain people, the U_{II} user shares the world with and perceives the substantial D_{II} resource produced by the A_{II} author, the U_{III} user conceives the formal D_{III} document produced by the A_{III} author, the U_{IV} user interprets the representational resource produced by the A_{IV} author, the U_V user receives and interprets the interpersonal resource sent by the A_V author, and the U_{VI} user is affected by the effective resource created by the A_{VI} author. Users in turn author queries, producing queries of the various facets (fig. 2.9).

A *user* was defined above as the association of relationships between the query, the ideal resource and the context that surrounds and is defined by the reader-cum-searcher (see section 1.1.3). However, the six facets also apply to these relationships, making a user even more complex than suggested above; users are faceted because the world that they inhabit is faceted. They have intentions and inhabit a world consisting of a network of intentions (U_I); they have substance and must obey the laws of the physical world (U_{II}); they have form and meaning and interact with objects that have form and meaning too (U_{III}); they have value and interact with things that have symbolic meaning, such as language (U_{IV}); they inhabit a world of situations and problems that must be solved by using objects and taking rational decisions (U_V); and they have feelings and emotions and are affected by their environment in various complex ways (U_{VI}). All of this shapes them as human beings and also guides their behaviour, in search as in other activities. Consequently, the facets are crucial for an understanding of relevance assessment in particular, which will be discussed thoroughly in Part II. Furthermore, they represent different stages of reconstruction that the system designer may attempt to reach. Having access to a D_{IV} document, it may be possible to reconstruct what the A_{IV} author intended to portray, but not the message that the A_V author intended to convey. Similarly, a Q_{IV} query, i.e., a query enriched with semantic information, may reveal what topic or other semantic information the portraying user is searching for, but it will not allow the system to reconstruct the U_V user with a need to apply this semantic content in real life. If this is wanted, a Q_V query is going to be needed in which contextual information on the user's application needs are provided. Richer queries simply imply a need for more contextual information.

Interpreting a piece of language is a highly complex affair due to the various facets inherent in such a resource. Imagine a simple resource with no linguistic content. The U_{II} user perceives it, the U_{III} user conceptualizes it and builds a mental construct of the various observable features, the U_{IV} user interprets the observed and is aware of the ontological relations involved, and the U_V user attempts to make sense of it and understand the whys and hows. Unlike the interpretation process, there is no clear order of the facets in production; intention pervades the process, and the facets are constructed more or less simultaneously.

A D_{III} resource cannot be perceived. True, the linguistic content is coded in orthography, but crucially the formal linguistic resource is the product of a mental process in the reader. This can only be interpreted by U_{III}, which is not a perceiving but an understanding reader. Similarly, the representational D_{IV} resource is an interpretation of the semantic content constructed by the A_{IV} author and so can only be interpreted by U_{IV}. Finally, the interpersonal resource is a message conveyed to the reader and hence can only be interpreted by U_{V}. The layout of a resource can of course be perceived objectively by U_{III} as part of the formal resource, but it only affects the U_{VI} reader psychologically. All of the dimensions are tightly bound together and intertwined in this way, but they act separately and have separate values. For instance, when you realize that the substantial resource is something you can handle, the formal something you can perceive, the representational something you can understand the general use of, and the interpersonal something that can be useful in a particular situation, it is clear that the relationships between the facets are very intimate. On the other hand, it is also clear that you cannot handle a linguistic text in abstraction from its physical document, understand a message without reference to its semantics, or use the psychological effects of your own reading of the document to solve problems without a situation in which to apply them[37]. The precise relationships between the various facets and dimensions will become much clearer with the introduction of the Regulated Flux model in Part II, which is named for this very ability to describe the flux between dimensions and the way it is nonetheless regulated by natural processes.

2.3.3 Situations

Interpreting a stimulus is also what a user does when faced with a task or situation that leads to a search. Consequently, situations can be analysed according to the Faceted Stimulus model in the same way as resources (fig. 2.10).

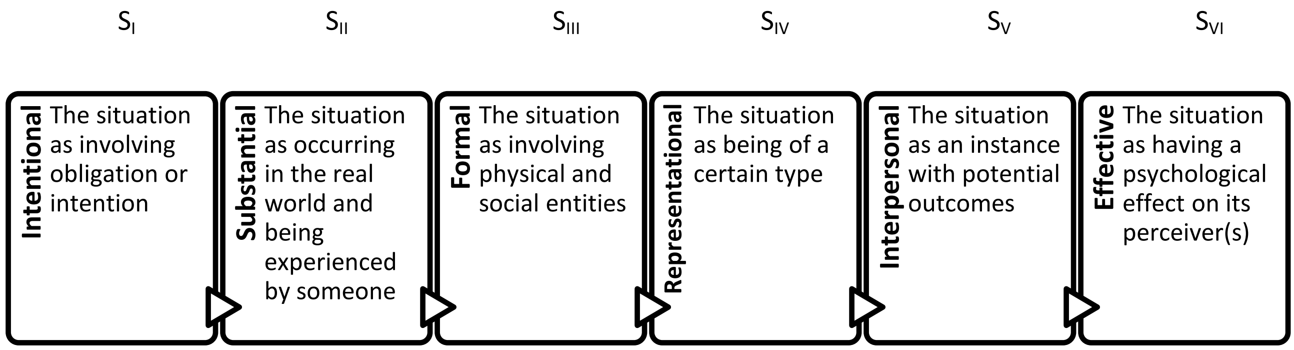


Fig. 2.10. The facets of situations.

The physical perceiving U_{II} user lives in the external world and experiences the problematic situation, the U_{III} user constructs a mental representation of the perceived context which can be understood on

[37] You can of course use the psychological effects to solve a problem, if the “problem” is itself of a psychological kind, such as boredom or a desire for aesthetic content, but you still need a mental context in which to apply the effect.

this level due to human embodiment (Lakoff, 1987), the U_{IV} user is able to understand it as being of a certain type, and thus to describe it, and the U_V user may reason and recognize what is needed in that particular instance, and thus form an intention to search for information. Finally, U_{VI} is affected psychologically by the situation. By treating the effective situation with its psychological effects as additional to the interpersonal situation that can be rationally evaluated as regards possible actions and outcomes, the Faceted Stimulus model mirrors the fact that psychological effects are not naturally covered by rational relevance evaluations unless the user is specifically aware of them.

It is this highly complex input that forms the basis of the developing need. It is no wonder, therefore, that understanding need construction is not entirely straightforward. Even if people were to submit full descriptions of the problematic situation facing them, interpreting such data would present severe difficulties. Firstly, the real situation causing the need is perceived by the user in a way which may not mirror reality accurately. Hence, users may not have interpreted their task correctly, and often revise their perception of it as part of the search process. The user's perception of reality is interpreted by the user in a way that incorporates emotions and background knowledge (and possibly distorts the perceived reality). The situation may then be described by the user through language, which causes the construction of a mental world (see chapter 5). This is subsequently substantiated in an actual linguistic description, a faceted query. Finally, the description is decoded and interpreted by the system or an intermediary trying to help the user. There is more than enough room for misinterpretation in this process.

2.3.4 Tasks and causes

When confronted with a task, people may approach the internet with the view to finding a solution. In such situations, they very often enlist the help of a LAT in order to find what they need, either by using web search or searching a particular site. Users of search engines phrase their queries differently depending on the kind of need they have and the kind of result they wish to achieve. Sometimes, the incentive to approach a LAT is pure **serendipity** – people will search or browse for mere entertainment, without any specific goal[38]. More frequently, however, they consult a LAT because they have a more or less definite **task** to complete (Ingwersen & Järvelin, 2005a). In fact, even serendipity may be considered a task in that it represents the purpose of being entertained, and this goal may be achieved or not. The fact that people are initially often very vague and uncertain about what their tasks

[38] Depending on the definition of this term, it may or may not include cases where the user's motivation is simply a desire to be informed in general or aesthetically stimulated, without any specific preferences; see section 8.3.

are does not mean that they do not have one; a task is not a clearly formulated question but merely a situation that requires the user to find some information.

Tasks appear in different guises. Lexicographers make a basic distinction among **cognitive**, **operative**, **communicative** and **interpretive** tasks (Tarp, 2008; Tarp, 2009; Bergenholtz & Gouws, 2010)[39]. According to this state-of-the-art typology, users facing a *cognitive* task seek information with the sole purpose of expanding their knowledge, whereas those with an *operative* task have a system-external problem that they wish to solve. Often, information may contribute to the fulfilment of such a task. The difference is that in a cognitive situation the whole point is to acquire information, whereas with an operative task the information is gathered with the solution of a certain non-informational problem in mind. *Communicative* tasks are those that present translation or understanding difficulties, i.e. text production, reception and translation. Dictionaries are typically used for such purposes, but interestingly enough search engines such as Google are increasingly approached in this kind of situation as well (Bergenholtz & Gouws, 2010). Finally, *interpretive* tasks concern the interpretation of non-verbal signs. This typology of tasks clearly reflects the interests of lexicographers in that these are certainly the main types of tasks that confront users who decide to approach a lexicographic tool such as a dictionary or an encyclopaedia. As our definition of *tool* is broader, the types of task associated with such tools are correspondingly more varied, but the basis is similar, and there appears no longer to be any clear borderline between the territories of lexicography and IR research as both fields open up and generalize their work.

By definition, users of a LAT communicate their needs by linguistic means. It is the aim of this thesis to attempt an analysis of such expressions in order to enhance search engines' and similar tools' "understanding" of user needs. However, linguistic analysis is hardly possible in the case of communicative and interpretive tasks because here the expression has simply not been constructed by the user but rather by the original author or sender of the expression or sign. We shall hence concentrate on cognitive and operative tasks.

Just as users have various reasons for approaching a LAT, the reasons that authors have for wishing to inform readers are also diverse. The broadcasting of information generally derives from some **cause** – whether for the direct or indirect benefit of the author, the reader or some other party. We shall reserve the term *task* for referring to the situation that caused the user to have an intention of

[39] The terminology within lexicography has still not stabilized entirely. The *operative* function is thus also known as *operational*, and the *interpretive* as *interpretative*. Lexicography is a field in rapid development, and some of the types have only very recently been recognized. Consequently, older sources only distinguish between two or three types of task (hence the various sources, which provide different typologies). Note that there are many subtypes.

searching, while *cause* shall be used as an expression for the situation that caused the author or selector to have an intention of informing. As authors are not present in the search process, they might be thought to be irrelevant in this connection, but in fact what made someone produce a resource is crucial when it comes to the question of who it is relevant for. If the user is not in the target group, the resource in question may be useless, but if, on the other hand, the user is in exactly the kind of situation that the author had in mind when writing a document that would inform or help such a person in some way or other, then that document is more likely than not to be among the most useful inputs in the results list, especially when dealing with advice. The following cause types are common.

- **Service causes**

  The author wishes to help the reader. Examples are citizen centres, health centres, etc. The author's willingness to help certain kinds of users opens up the possibility of drawing on their goodwill for tagging resources according to the situations in which they might be useful, and to whom.

  As noted in the section on methodology, the user is a function of context. The function is basically one of applying a perspective to the situation surrounding the user, both physically and psychologically. The situation that needs to be conveyed to the system is this perspectivized version, i.e. it is actually a description of the user as regards the most relevant features in the current circumstances. The effect of this is that tagging a document with a type of situation is equivalent to tagging it with a user type.

- **Altruistic causes**

  The author wishes to convince the reader of something for the benefit of some external cause, which may or may not concern the user personally. Examples are campaigns for environmental, health-related, political, or other causes, but also scientific papers and information leaflets, etc. As with service causes, such authors are highly interested in having their texts read, which allows us to rely on their willingness to tag documents, though there is often no question of usefulness in actual situations.

- **Promotional causes**

  The author wishes to convince the reader of something for the author's own benefit. Examples are advertisements for products and places to visit, and campaigns for certain types of causes that benefit the author (or the author's organization) directly or indirectly. With regard to the author's willingness to tag resources, the same applies to promotional as to altruistic causes.

● **Recommendatory causes**

The author wishes to recommend something, or to dissuade others from buying an item or visiting a place, etc. Once again, authors may be willing to tag their resources according to the kind of (psychological) situation in which users might be tempted to access a certain external resource.

● **Social causes**

The document was written for social or entertainment purposes. This often involves self-promotion without obvious direct benefits.

● **Reporting causes**

The author is reporting an event or a statement made by someone else. Specifically, selectors may not have any intention of conveying a message except reporting what the author has written verbatim.

● **Artistic causes**

Fiction comes under this heading.

Since the central goal of a search engine is here assumed to be to help the user, we shall concentrate on documents resulting from *service*, *altruistic*, and *recommendatory* causes.

2.3.5 Dealing with facets and user behaviour

When using web search and most other LATs with unordered data, users have no way of knowing what documents are out there[40], and consequently cannot know what the authors' intentions might be before seeing the documents; they are effectively expected to reconstruct potential intentions behind possible relevant documents before coming up with a suitable query. Searching for a substantial resource that nobody had the intention of creating is pointless, as is searching for a linguistic document that was not intended to be a form of language. A user searching for a certain text string would hardly be helped if a document were retrieved which contained that exact string, if it were only through an accidental arrangement of random characters, and searching for a representational document where nobody had the intention of making a portrayal would likewise be meaningless. If you search for a sentence and retrieve a document that contains this exact sentence, but it turns out that the document (and the sentence) is in a different language from what you expected, and that consequently, the semantic content of the sentence is not what you had anticipated, the document is simply not (likely to

[40] Some cases of enterprise search and searches performed in closed small document sets may be exceptions, but otherwise the set of potential resources is basically unknown.

be) relevant. Finally, when searching for an interpersonal document, receiving one in which nobody intended to convey that particular message is utterly unhelpful. If your search for *Copenhagen cafés* were to result in your finding a fictional text describing cafés in Copenhagen, you would probably not be satisfied, because you were more likely to be interested in a list of real cafés and not a fictional description of possibly fictive cafés. Consequently, what we need is for the LAT to be able to interpret these different layers of intentions for the user and retrieve only those documents that really do convey the intentions relevant to the search. What layer the user is interested in is something that he or she will have to convey to the LAT in some way or other, though in some cases it may be inferred from context with varying degrees of confidence.

Identifying the user's intention is useless unless documents can be tagged in some way according to what intentions they match. Tags can be supplied to all resource facets. With linguistic documents, the substantial facet of a resource may be tagged with the script in which it is written, the formal facet with the language in which it is expressed, and semantic notation might be used for the representational facet. In a similar manner, one might tag the interpersonal facet of a resource according to the intention of the informer. One way of doing so would be by supplying possible situations in which users might find the document helpful (Blair, 2003).

> The first step in information retrieval system design, then, is to develop a detailed taxonomy of the various activities and practices that produce or use the information on the system. Each document on the system must then be explicitly linked to one or more of these activities or practices. For private industry, documents could be linked specifically to the value-creating activities and the core competencies of the firm. (Blair, 2003, p. 38)

While the simplest form of indexing simply consists in listing the words and phrases that occur in a document, much more can be done. Additional information may be added during or after crawling, e.g. by parsing the documents in an attempt to understand their contents. Semantic tags may be added which disambiguate and clarify the meaning of the words and sentences. In a similar manner, it would be possible to tag a document with the intended uses of the document, or the way in which it might be helpful, and to whom.

There has been a growing interest in formal representations of the semantic contents of documents. The idea from the point of view of search is that a semantic description would allow search engines to match queries not only to the substantial documents but also to their semantic contents, thus matching queries to what the documents are really about (Riad, Eliminir, ElSoud, & Sabbeh, 2010). This is certainly an improvement, and semantic search is in any case extremely useful, but there is a risk of falling into the trap of thinking that semantic representations will help us provide truly *useful* resources when in fact we cannot rely on the linguistic contents of documents to identify the author's intentions but only the representational facet of resources which contains their semantics. Firstly, fictional texts and other reports mask the real intention behind a document, which can be pure entertainment or information with quite different aims from the original source. Also, the intention may decide how the contents are to be interpreted; a text in which a politician writes about the conditions in Danish schools may not in fact be very informative about the conditions of the schools, but only about the way politicians should go about improving these same conditions, in spite of containing many references to schools. Consider a statement such as the following.

2.6	There are very few schools in Northern Jutland.

Taken at face value, this might cause you to decide that finding a school for your child in Northern Jutland is hopeless. However, this would be an erroneous conclusion, because the expression *very few* should be seen in relation to the political intention, and thus has nothing to do with a situation in which people are looking for a school, since for any one child there may in fact be many schools to choose from.

By letting informers tag documents with their intentions, users' intentions can be matched with those of informers, while at the same time excluding from consideration those documents which were composed without any real intention of informing proceeding from a service, beneficial, or recommendatory cause. This will make it possible to distinguish between documents that merely mention a topic and those that actively seek to help users in specific situations, which will most likely be of great advantage to users. However, some users do target other facets, such as the representational facet as in prototypical topical search, and we shall in no way confine ourselves to cases where information is needed for a clearly articulated purpose or published with a clear cause in mind.

Note, finally, that users should preferably always be given the opportunity to revert to a simpler type of search if they are actually interested in a "simpler" facet of the document. One should not be led into thinking that the most sophisticated system that is able to read between the lines and return information rather than text is invariably most helpful to the user, even if this is probably true in *most* cases; sometimes, users do perform simple string searches and at such times they prefer to get exactly what they ask for.

2.4 Search as communication

Ambiguity of expression is a major issue in IR research because queries tend to be short and vague. Yet, according to Relevance Theory, true linguistic ambiguities should not occur, since speakers (in our case, users) are expected to want their listeners to understand them and hence to formulate their messages in a way so as to be easy to understand.

> A speaker who wants her utterance to be as easy as possible to understand should formulate it (within the limits of her abilities and preferences) so that the first interpretation to satisfy the hearer's expectation of relevance is the one she intended to convey. (Wilson & Sperber, 2004, p. 259)

This cooperative behaviour should preclude ambiguity, because an ambiguous message would cause the addressee unnecessary interpreting effort and hence would not qualify as an optimally relevant message.

Why then do IR researchers worry about ambiguities?

The reason is that though queries are unambiguous when interpreted in context, their precise meaning cannot be determined when they are treated in isolation, and unless contextual information is explicitly collected, a query will be isolated from the system's point of view. Even though users are not aware of it, they leave out an enormous amount of information that is necessary for a correct interpretation of their request. Some interpretations may be considerably more likely than others, but this is not always the case; sometimes queries are simply ambiguous. Consequently, it is a major goal of IR research to attempt a formalization of context in order to get at the part of the intention which is left unstated.

Although there are clearly some differences between search and more prototypical discourse the similarities are substantial, and this is important, because it allows us to draw on the work in discourse studies when trying to understand queries, and also subsequently to do the opposite. For an illustration of the similarities, imagine a person in need of some information and compare a scenario in which that person chooses to ask a friend for help to one in which he or she decides to consult a LAT instead. The following steps must be gone through in the two cases.

● Getting answers through discourse

A speaker who wants to ask a question or make a request of someone goes through a series of stages from recognising a need for information to receiving an answer. The stages can be summarized as follows.

1. The speaker faces a task

- The speaker recognises a need
- The speaker approaches an interlocutor for help
- The speaker constructs a linguistic version of his or her need, adapted to the requirements of the interlocutor
- The interaction with the interlocutor may influence the speaker's cognitive state
- The speaker may revise his or her needs based on a change in the cognitive state

2. The speaker receives a response from the interlocutor

- The response influences the speaker's cognitive state
- The speaker may revise his or her needs based on the change in the cognitive state
- The speaker judges whether the response was relevant, i.e. whether the needs have been answered

● Getting answers through information retrieval

A LAT is likewise an interlocutor from whom answers and information can be retrieved. The stages that must be gone through in this case are very similar, although the interlocutor is admittedly somewhat different, as is the kind of expected response. The main dissimilarity is that responses are often initially in the form of **results lists**, which the user then has to assess before accessing an actual resource, which is the ultimate response to the request. The decision about what resources seem to be relevant is normally based on a short extract from the resource called a **snippet**. Here is a schema of the search process[41].

[41] Some LATs do not return lists of references, and some do only that, so that either Stage 2 or 3 may be left out depending on the kind of tool.

1. The user faces a task

- The user recognises a need
- The user approaches a LAT for help
- The user constructs a linguistic version of his or her need, adapted to the requirements of the LAT
- The interaction with the LAT may influence the user's cognitive state
- The user may revise his or her needs based on a change in the cognitive state

2. The user receives a list of suggested resources

- Reading the list influences the user's cognitive state
- The user may revise his or her needs based on the change in the cognitive state
- The user assesses the likely relevance of documents based on the information in the snippets

3. The user accesses a resource

- Reading the resource influences the user's cognitive state
- The user may revise his or her needs based on the change in the cognitive state
- The user judges whether the resource was relevant, i.e. whether the needs have been answered

The two processes are clearly parallel. The main phases of search are known as the *problem state*, *system interaction*, and *document interaction*, respectively (Bruce, 1994).

Bruce (1994)

- **Problem state**: The problem state comprises the problem, internal knowledge state, intent, and public knowledge estimate (Saracevic, Kantor, Chamis, & Trivison, 1988) of the user[42]. It is a state of readiness that implies that the user's knowledge structures are sufficient for a preconceived notion of what will be considered relevant to the problem state and what will not.

[42] The latter implies an estimate of what relevant information is likely to be available at all.

- **System interaction**: Search tactics relate to the problem state of the user and are altered or validated by responses obtained from the information system engaged. The user's interaction with the system is considered a learning process resulting in some development, alteration, enhancement or modification of the problem state.
- **Document interaction**: The point of access to and interaction with the document(s) in the document store (library).

Note that our sketch of the search process as a parallel phenomenon to discourse in general shows what goes on at the discourse level only; the process is far more complicated from the point of view of users, mainly because the question that they need to ask is often not clear to them from the start[43]. This may of course also be true in some ordinary discourse situations, but it is especially characteristic of search.

If the same cognitive processes are active in search as in communication in general, then it follows naturally that users may unconsciously require a LAT to infer the same kinds of information from a query as human listeners would. People are likely to adapt some features of communication to the fact that their interlocutor is a computer because in general they expect computers to be unintelligent and inattentive to context, but the process must still *in essentials* be the same as in discourse in general, and it is suggested here that some features of communication are simply so deeply ingrained in people's minds that they never stop to think that they might need adjusting when the addressee is not human. The image of computers as unintelligent may be changing with improving technology as people grow used to new, sophisticated features, but as yet there is certainly a great deal of truth in the presumption that computers are linguistically incompetent. Even so, we can learn important lessons about the search process from the study of discourse.

The modern view of discourse was founded by Grice (1975) when he introduced what has come to be known as the *inferential model* used to understand it. The classical view before Grice came along was that an utterance encoded a message, which the listener simply needed to decode. Listeners would be able to do this because they spoke the same language as the speaker and thus possessed a copy of the code and could apply it in reverse. With the work of Grice, this all changed. He appreciated that

[43] In a questionnaire study, Saracevic and Kantor (1988) found that 58% of the participating users believed that they had their problem clearly defined from the start. The actual percentage is likely to be considerably lower, depending on what one understands by "clearly defined".

messages do not in fact contain in their codes all the information that they convey. Rather, much of the meaning of an utterance has to be drawn from its context and is left to the listener's powers of inference. Grice posed himself the intriguing question how it is possible to communicate ideas without specifically putting them into words, and he developed a theory based on the understanding that communication is a cooperative effort; it is in the speaker's best interest that the listener should understand the message, and so messages are always framed in a way that allows them to be decoded in context. A speaker draws on context for simplification when coding a message and expects the hearer to be able to do the same for meaning expansion in the interpretation phase. Some things may be left out, either because they are obvious when taken in context, or because the hearer can be expected to be able to infer the meaning from a combination of the linguistic cue and the context. For a simple example, imagine an American going to a ticket counter and uttering the following sentence.

2.7	I would like to watch a football game tomorrow.

There would not normally be any need for the speaker to specify whether European or American football was meant; this would be clear from context. Also, the clerk behind the counter would be likely to interpret the utterance as a request for 1) information on football games played on the following day (in an appropriate location), and 2) tickets for such a match even though none of these interpretations are encoded in the message; they simply arise from the combination of the expectations raised by someone uttering the sentence and the situation in which it occurs.

With queries such as the following, we are less fortunate as regards contextual clues unless we can access some additional information.

2.8	**football**	*
2.9	**football tickets**	*

A machine cannot at present be expected to be able to get at the implicit part of the message, which is so integral to it that it hardly makes any sense to exclude it. An IR system neither has access to the complete context, nor the ability to decide what parts of it are significant in a given situation. This is why retrieving information with reference to nothing apart from the query string is rarely sufficient; to a machine the string simply does not convey all that the user intended it to. The question is whether and how IR systems may be made to understand user intentions better by drawing on contextual information and linguistic clues.

We cannot expect users to avoid ambiguous terms, because that would mean requiring them to visualize their task in an ultimate, universal context so wide that their human minds cannot handle it; users may not even notice the fact that a certain term is ambiguous until they receive irrelevant responses from the LAT, as they are so immersed in their own immediate context that they often do not look beyond it. It is thus probable that few people even consider the fact that terms such as *apple* or *jaguar* are ambiguous the moment they use it in a query; they only realize this when the results are delivered, and of course the next time they use that particular word in a query they may have learnt a lesson and remember to disambiguate it in some way. Consequently, LATs need to be designed to cope with expressions that draw on context by collecting at least some contextual data.

Today, the impossibility of interpreting language without access to context is generally acknowledged (Dam & Dam-Jensen, 2007). However, because written language is so much more frequently analysed than spoken language, there is a focus on language *interpretation* in linguistic studies, when in fact the same is true of production; linguistic expressions cannot be properly produced without access to context, because meanings are simply too complex to be communicated without the reduction that context allows. It is thus likely that the simplification resulting from exploiting context is what makes communication possible in the first place; it would not be possible to convey any ideas without simplifying them immensely and letting the user imagine the rest[44]. It is of course not certain that the hearer will make the correct interpretation from the fragmented clues, but on the other hand, the speaker may have intended to be ambiguous on certain points.

Designing LATs to tune in on the current environment may be achieved either by understanding the user better in the long term by employing user profiles or by improving the understanding of the task at hand, and not least the discourse that the query partakes in. Chang and Lee (2000, p. 6) call for "a conceptual framework of context that can: (a) represent its multi-dimensional and dynamic nature; (b) demonstrate the relationship between context and information behavior in a systematical way; (c) generate [a] theoretical contribution based on present theories of information seeking and use". It is hoped that the following sections will be considered to contribute to the framework (a) that will allow (b) and (c) to be carried out.

[44] As pointed out by Fauconnier (1997, p. 190f), it is reasonable to suppose that animals too are able to construct the mental models of a situation that represent the first step in portraying it (the first point of subjectivity in the bisubjective model introduced below). What makes humans stand apart may thus simply be our ability to reduce the enormous complexity and mass of knowledge connected with such cognitive models into something that is compact enough to be transmittable through language.

2.4.1 Context in IR

Traditional search engines provide the same result in response to identical queries issued by different users in different situations, but at a certain point IR researchers started noticing that user needs may vary extensively with context and decided to try to incorporate contextual parameters into search systems. This has been done in several steps, and different traditions include different kinds of context. As a preliminary analysis, and for the purpose of giving a brief introduction to the development of context models for IR, search context may be divided roughly into four main spheres as follows (fig. 2.11).

1. **The external user context**: The user's interaction with the external world, i.e. the task confronting the user, and his or her personal background and preferences.
2. **The internal user context**: The user's interaction with the system, i.e. the query, previous queries, the site from which the tool was accessed, etc.[45]
3. **The internal resource context**: Features of the actual documents, such as topics treated.
4. **The external resource context**: Contextual information on the author of the documents and/or possibly on a selector who published them; the task (for which we have chosen the term **cause**) facing the author and prompting the production of the document, as well as information on the author's background.

All of these main types are highly complex and many-layered, not least because of the facets which permeate all of them.

The very notion of context implies an assumption that there is somehow something at the centre to which everything else is contextual. The question is rarely posed what this central notion is when dealing with search. From the system's point of view it could be the query, and this is probably the interpretation that most IR researchers would make, yet from the user's standpoint it is possible to see the actual aims and goals of approaching the LAT as at least being part of the centre to which everything else is context. Alternatively, the user's task could be taken to be central, making user

[45] The word *internal* should here be taken to mean *system-internal*. It does not mean psychological as opposed to what happens around the user.

characteristics and real world features related to the task contextual, or the desired documents could be central, making the use to which they might potentially be put contextual. Strictly speaking, few features of these contextual spheres are closely contextual in relation to the query, which has an extremely shallow context as such – in its character of a search string. Yet because it is bound to user tasks as well as resources, it connects with these other contextual spheres and incorporates their features depending on the "centre" chosen. Consequently, the context of a query in this broad sense is extremely complex, combining user, document and query context in a narrow sense.

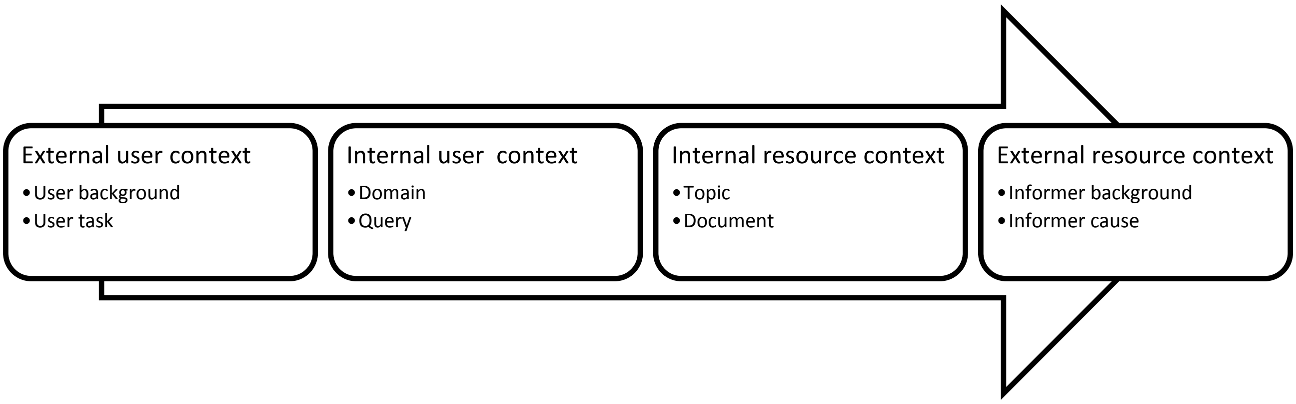


Fig. 2.11. The four main contextual spheres as proposed here. Note that the domain to which query terms belong parallels the topic to which document terms contribute and therefore does not belong with the external task.

The first researchers to include context in their understanding of the search process focused on the internal user context, and in particular the domain within which the search was performed. It was found that ambiguous queries can be disambiguated by taking such context into account as users tend to be more interested in certain knowledge areas. Thus, a biologist would be likely to refer to the animal, the plants and the island when submitting queries such as *python*, *palm*, *apple* and *java*, whereas software developers would presumably see such words differently. These techniques were known as **adaptive IR**. Later, they evolved into **contextual IR**, which "aims to enhance the understanding of human needs, activities and intentions in order to deliver more accurate results in response to user queries" (Tamine-Lechani, Boughanem, & Daoud, 2010, p. 3). Consequently, the notion of context was broadened to also include the external user context and the internal resource context. Some approaches focus on the former, successfully returning different results according to user task. This is the lexicographic solution, but it is also the approach taken by researchers working in context-aware IR in mobile applications and ubiquitous computing, for the simple reason that the user's geographical position is often crucial in determining relevance in such applications (Martins, 2009). Other types of users (and hence, researchers) are more interested in features of the resources, such as the geographical and temporal origin of the document, or the type of document. This is usual in searching for weather information, for instance (Barry, 1994).

Contextual IR led in turn to a strong focus on **personalized IR**, which normally includes external and internal user context only. It is a further development of adaptive IR in which the user task is taken into account. By employing user profiles and registering user behaviour, relevance may be predicted to a certain extent, though such solutions tend to overestimate the helpfulness of finding results that match the user's (general) interests and underestimate the distinctiveness of different queries with the result that some such systems described in the literature will find more or less the same results no matter what the user is searching for[46]. This is of course a matter of the correct tuning of the weights of general interests and actual query topics. In spite of its drawbacks, personalized IR is perhaps the most popular approach today, not least in the branch fuelled by commercial interests, and the notion covers a wide range of solutions, some of which are successful in improving search significantly, especially if applied in the right circumstances.

Two literatures have sprung up based on the *systems view* and the *situational* (or *user-oriented*) *view* of relevance, respectively. The systems view considers the search process with little reference to the user's situation and task, whereas advocates of the situational view care passionately about user situations but tend to forget about the search process as an interaction between a user and an IR system. The discussion of context in IR research has centred on transferring the main focus from the system to the user. This has led to many fruitful insights, but there is one dimension which has tended to be forgotten, viz. the provider of the information, or what was called above the *external resource context*. This is a serious oversight, because the external resource context is an extremely useful context type, as resources can be tagged for the author's intention to inform readers but not for the user's actual use of them (except perhaps through social tagging). The author of a document usually foresees the uses that the text can be put to as part of the production process. Admittedly, there will always be uses that were not anticipated, but documents are generally most useful for the purposes that they were made for. Therefore, the user will benefit from having documents that were made for a suitable purpose ranked higher. Of course, access to information on what the user actually needs is still required, so the external user context is also important. **Cognitive IR** aims to include all four main types of context (Ingwersen & Järvelin, 2005a, p. 191ff), and so this is the approach recommended here (fig. 2.12).

[46] See Phinitkar and Sophatsathit (2010) and Kajaba and Navrat (2009) for a couple of examples.

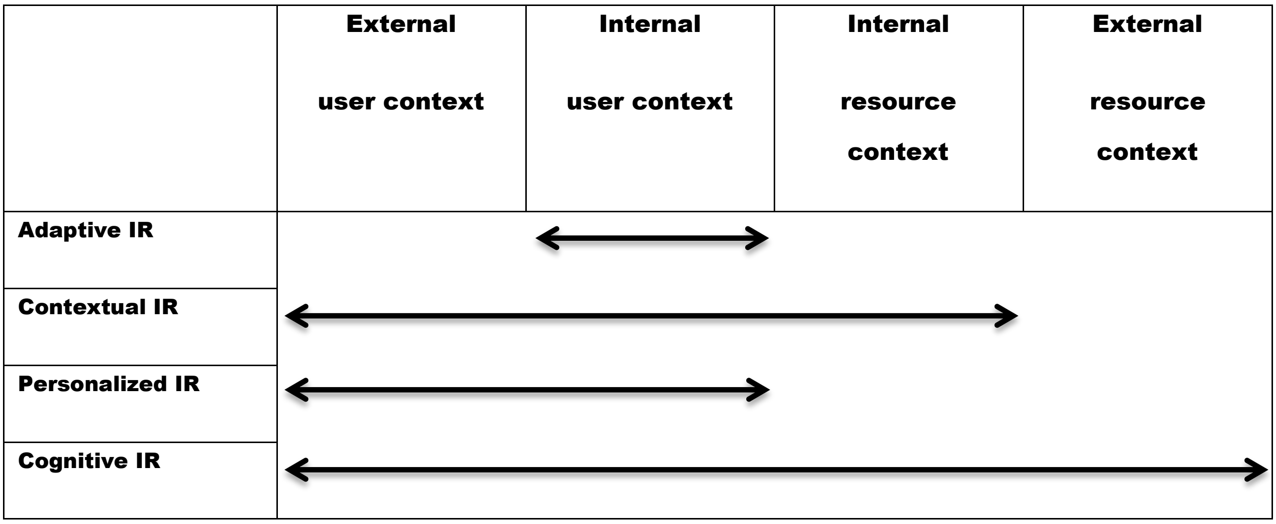


Fig. 2.12. The alternating foci of IR research, with the approach adhered to here shown at the bottom.

Collecting information on the user and context may help in several ways (Tamine-Lechani, Boughanem, & Daoud, 2010). Firstly, short queries may be disambiguated if the context tells us that one interpretation is more likely. This is exceedingly useful since most queries are rather shorter than is necessary for a satisfactory interpretation. In addition, retrieval precision can be improved if we know the user's main interests and preferences. Also, by ignoring information irrelevant to users interested in a certain domain only, users operating within a defined domain can be helped very effectively. Some systems moreover increase recall by searching for related topics as well as the one implied directly by the query. However, this is not always as simple as it might sound, because what concepts are to be considered related may depend on context. Even in such seemingly straightforward cases as subsumption, there are differences between domains. Thus, if you search for *pets* in a pet shop website, documents on rats are likely to be relevant, while if the same search were performed on the site of the local authorities, it would be preferable not to retrieve documents on rats, because they would be more likely to deal with pest control, which is not about pets at all. Finally, some contextual search tools encourage social interaction by letting users recommend resources and then comparing like-minded users.

The basic architecture of a contextual or personalized IR system consists of a **context-modelling module** which combines the query and other contextual data into a model of the user need, and a **retrieval module** which compares this model with the resources available and delivers a result. These two modules correspond to the two basic tasks of a search engine: Compare the query to the documents, and compare the document topics to the user interests. In practice, this may work in very different ways. Thus, improving search quality can be done at three points in the process: By **pre-**

processing the query, by **changing the core search algorithm**, or by **post-processing** the search results. Data from user profiles may be exploited at all of these stages (Daoud M. , 2009). For some reason, user context is not usually brought into the core as part of the query (Tamine-Lechani, Boughanem, & Daoud, 2010); most attempts at personalizing search use either pre-processing in the form of **query expansion** (in a narrow sense), or post-processing, also known as **rank biasing**[47]**.** While pre-processing aims at retrieving different and more precise results by narrowing the query, post-processing simply favours the best results by ranking them higher.

The question of the incorporation of context has mainly been researched with two different purposes in mind. Research in context-aware IR in mobile environments and ubiquitous computing concentrates on the effects of the spatiotemporal location of the user, as well as of the device from which the search is performed, while researchers in personalized IR attempt to incorporate implicit or explicit user representations into the search process, mainly through the exploitation of user profiles. Such profiles are typically based on implicit information inferred from the user's click history, eye-tracking, past page views, bookmarks or even desktop information, while further direct and indirect information on user interests may be gathered from news sources, blog sites and web shops (Tamine-Lechani, Boughanem, & Daoud, 2010). An alternative approach is to exploit **collaborative filtering** and so favour documents preferred by (seemingly) like-minded users. The information needed for any contextual approach may be collected **proactively** as click-through behaviour and previous search activity, or **reactively** in the form of direct feedback from users (Lee, Kim, & Lee, 2010; Tamine-Lechani, Boughanem, & Daoud, 2010). Whatever the case, this information needs to be stored in a database (or alternatively in the actual documents); this works well in narrow domains such as movie and book databases where the document collection contains a manageable set of resources (Park & Pennock, 2007). In a web search context, there are simply too many documents for this method to be practicable, since all documents would have to be marked for preferences according to all user groups. Also, defining like-mindedness is relatively easy in a narrow domain, whereas it would be extremely difficult to define user groups across a wide range of domains.

Personalization systems that depend on query expansion enrich the query with contextual information before comparing it to the resource collection. In its simplest form, this means adding extra words to the query, so that the query *python* may be interpreted by the system as *python snake* if the user is

[47] One exception to this is the personalized Bayesian retrieval model in which a keyword vector representing the user's topics of interest is integrated in the relevance calculation (Daoud, Tamine-Lechani, & Boughanem, 2009).

assumed to be more likely to be interested in snakes than software development[48]. The system then compares the query to the user profile in order to determine what concepts are most likely to interest the user and consequently disambiguate any ambiguous terms. One situation where such query reformulation is exceedingly useful is context-aware IR for mobile environments, which employ so-called *local search*: Users searching for *café* in a mobile environment are likely to be most interested in cafés nearby. The query may accordingly be expanded by adding the name of the town or another nearby location (Hattori, Tezuka, & Tanaka, 2007; O'Brien, Luo, Abou-Assaleh, Weizheng, & Li, 2009). In other types of IR system, query expansion can be extremely risky as users lose control of what they are searching for. Some systems actually seem to focus more on user profiles than the actual queries, which is obviously not very helpful, as it goes against the user's explicit wishes. If this method is used, care should at least be taken that expansion terms are assigned lower weight than the original query terms.

Even when it comes to presenting personalized or contextualized results to the user there is a variety of available solutions to choose from. In *rank biasing systems*, the query is left as given by the user, but the results are re-ranked according to the user's preferences and interests. In this way, the knowledge that a certain user is a biologist may be utilized in order to place the snake results at the top when the query *python* is provided, though the software results may still appear further down. This reduces the consequences resulting from making the wrong guess. A variant of this is *personalized categorization of search results* in which results are grouped into categories in order to facilitate finding relevant results without giving a definite opinion as to which category is the more likely to be relevant (Daoud, Tamine-Lechani, & Boughanem, 2009). This may in fact be a very helpful solution to the problem of identifying users' needs; simply let them choose, and you will not run the risk of making the wrong decision for them[49].

[48] Phinitkar and Sophatsathit (2010) is an example of this method, in which topical words assumed to represent the user's interests are added, based on ontological hyponymy relationships between topics so that subtopics register with main topics.

[49] An alternative to query expansion and rank biasing that should also be mentioned before closing this ultra-short review is the *meta-search* approach, in which the query is left untouched but many different alternative sub-queries are generated and sent to standard search engines. The results are then collected, combined and re-ranked (Glover, 2001; Kajaba, Navrat, & Chuda, 2009).

2.4.2 Defining context

Most definitions of context are extremely vague and non-committing, along the lines of the following.

> Context can be defined as all the circumstances, the data and information that are somehow relevant to the event or fact. Another definition says that it is the words or word phrases that give the sense of a word or phrase. (Kajaba & Navrat, 2009, p. 2)

A more precise definition is given by Tamine-Lechani *et al.* (2010, p. 3), who describe context as "any knowledge or elementary information characterizing the surrounding application (user, objects, interactions) and having an important relationship with the application itself". This definition seems to be based on the one provided by Dey (2001).

> Context is any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves. (Dey, 2001, p. 5)

In other words, everything that may have the slightest effect on the relevance of a resource is (relevant) context. Whenever we speak of context, it should be taken to mean *relevant context* in this sense.

As pointed out by Chang and Lee (2000, p. 5), context is sometimes considered to cover any features of the environment, while others restrict it to the immediate **situation**, where *situation* refers to "some specific conditions in a definite time-space movement, and is characterized as a set of related activities that occur over time"[50]. This definition would appear to basically restrict context to the task at hand. However, it also seems clear that many things that are not part of the task as such may nevertheless be relevant or even crucial to its fulfilment. It is apparent from their paper that what Chang and Lee actually meant was that a situation is simply one slice of context corresponding to a specific point in time. As such, a situation *is* context, and this means that all kinds of context may be included in the description of the situation as long as they are in some way relevant to the search. The immediate task is thus not the only interesting contextual parameter when designing a personalized tool. Rather, an overwhelming amount of contextual information on the user, the resource, the author, and the topics covered by the sought resources is of potential importance and could be collected.

[50] This formulation seems in turn to be based on Sonnenwald (1999): "A situation may be characterized as a set of related activities, or a set of related stories, that occur over time. That is, we can characterize or describe situations by actions or behavior that occur over time, and which are perceived as being connected by participants and/or outsiders".

However, deciding what is truly relevant is exceedingly difficult, and obtaining the information is not exactly easy either.

The idea that context is dependent upon the concept of relevance is inherent in the definition given of context for our purposes in section 1.1.3 stating that context arises around a query and consists of everything that is relevant to that query. In that definition, only user (or query) context was considered, and this is a common state of affairs. However, as mentioned above, author (or resource) context may also contribute with essential information. The inclusion of this other type of context requires a restatement of users for our purposes as $u(Q, C^Q, D)$ and of informers as $i(T, C^T, H)$, where C^Q is the context centred around the user issuing the query and C^T is the context around the resource-producing informer; obviously, neither can know the context of the other. Q is the query and D the desired document set, T the resource and H the targeted readers (fig. 2.13). Both types of context need to be included in a context-aware IR system in order to understand resources as well as queries before comparing them.

Static relation (Bound variable)	Free variable
The resource production $p(C^T,T)$	The targeted readers H
The message $s(T,H)$	The resource context C^T
The agenda $g(H,C^T)$	The resource T

Fig. 2.13. IR research setups for studies of resource production showing the relationships between authors (outside of the table but depending on C^T, T and H), resources and readers.

As pointed out by Chang and Lee (2000), there are conventions and conventionalized task procedures which precede the user's cognitive involvement and can thus be described without consulting the cognitive space of the user. This is certainly true, but the point is not that no part of context can be described without recourse to the user's mental world, but rather that no part of it is *only* accessible from outside the cognitive space. Hence, what the user is aware of may be considered the outmost limit of relevant context[51], unless one would like the system to be wiser than the user and risk responding in ways that the user does not understand. This latter view would be an entirely valid position to take, but it would mean that no limit at all could be imposed on relevant context, since the

[51] In the widest sense, i.e. the user need not be *consciously* aware of something for it to be part of context. Cf. Stephen Harter's notion of *knowing in a weak sense*: "to say that an individual's cognitive environment contains a given fact – that is, is manifest to him – merely means that the fact is capable of being perceived or inferred, not that it will be perceived, now or ever" (Harter, 1992, p. 604).

system would potentially need to have access to all kinds of knowledge relevant to the search context irrespective of the user's request. Some people will undoubtedly maintain that users are often unaware of many things that are in fact crucial to a successful search and that it is the job of a good search engine to suggest material that even the user hadn't realized was needed. This is a troubled discussion, because there are arguments for and against ignoring (or bending) the user's explicit request.

2.4.3 An improved typology of context for IR

The difficulties in designing a model of context are immense. Since in the end everything can be said to have at least some relevance to everything else (however remotely connected), there is no limit to what might potentially be included in a description of context. Obviously, most contextual facts would be extremely unhelpful in practice, but actually drawing the line is very difficult indeed. In practice, a rather narrow and to-the-point context description is probably more useful than an all-encompassing one where competing pieces of information weaken each other.

> The issue is not to apply as much context as possible, but to be selective and consistent, opting for realism at the same time as being in investigative control. (Ingwersen, 2007, p. 23)

The problem is deciding what focus to adopt. Designers of contextual IR systems typically concentrate on one of three types of contextual information: 1) users and their needs, 2) the goals and aims of users, and 3) the properties required of resources by the users (Daoud M. , 2009). In fact, these three context types may be considered as referring to the user's needs at the three main stages of the search identified above: 1) *The user faces a task*, 2) *the user approaches a LAT with a purpose in mind*, and 3) *the user evaluates the response with some criteria in mind*. This means that the relevance of a certain contextual fact – as well as of a given document – fluctuates during the search process. Indeed, the dynamicity of context is one of the most important obstacles in contextual IR research. An interesting point made by Sperber & Wilson (1986) is that context is not even given at a certain point in time. In discourse, hearers select the context that results in the best balance between contextual effects (i.e. cognitive effects derived from interpreting the utterance in that context) and processing effort. In the case of a LAT, therefore, the system should ideally choose the context by applying these same criteria. This is of course rather difficult at present.

The following are the most widely discussed contextual dimensions mentioned in the literature (Tamine-Lechani, Boughanem, & Daoud, 2010).

- What kind of **device** or platform is used for the search?
- What is the **spatiotemporal context** of the search?

This is highly useful if resources are not valid across time and in different geographical locations, such as weather information.

- The **user context**
  a. The **personal context**
     i. **Cognitive context** such as the user's level of expertise and main interests and preferences
     ii. **Psychological context** such as anxiety and frustration
     iii. **Demographic context** such as language and gender
  b. The **social context**, such as friends, colleagues and like-minded users

Recommender systems based on collaborative filtering techniques exploit the idea that social context determines some of the user's preferences.

- The **task**, **problem** or (more generally) **situation** behind the information need

This gives rise to different goals and aims which affect the way queries are expressed.

- What is the **document context** in which the information is embedded?

However, the fact that they are widely used does not necessarily mean that they represent the ideal set. In fact, a complete list of relevant contextual dimensions would require the addition of many others which are habitually overlooked. A few researchers have taken the time to sit down and figure out what kinds of context might theoretically be relevant to search, but every such typology is different due to the fact that various foci are chosen by different authors depending on the purpose of their work, and in addition the various distinguishing criteria are assigned different amounts of weight. Based on the literature on context, Chang and Lee (2000) arrive at the following main contextual dimensions to be considered in IR[52].

[52] The sources used by Chang and Lee are Wilson (1981; 1997), Taylor (1993), Rosenbaum (1993), Marchionini (1995), Cool (1997) and Sonnenwald (1999).

1. **Access**: Type of access and procedures
2. **Cognition**: Person, people, psychological ecology, cognitive orientation
3. **Outcome**: Retrieval outcome
4. **Problem & resolution**: Person with information problem
5. **Resources**: Information resources environment, information horizon
6. **Roles**: Social roles
7. **Rules**: Organizational structure and procedures, conventions
8. **Setting**: Environment, physical setting, economics constraints
9. **Situation**: Interaction situation
10. **Task**: Task at hand, task environment

These types do not form any transparent pattern that would make it possible to easily check whether they are indeed the necessary and sufficient types of context for the purposes of IR. However, it is clear that there are several layers of relevant situations involved: The interaction with the tool itself (*Situation*), the problem solving situation or information need satisfaction related to the problem that prompted the search (*Problem & resolution*), the overall task that made the user approach the tool in the first place (*Task*), and the wider setting of which the task is a part (*Setting*). To this should be added the individual cognitive and collective social spheres, the self and the community and the relationships between people (*Cognition*, *Roles* and *Rules*). Furthermore, the device on which the tool is used and the resources accessed (*Access*), and the resource collection searched (*Resources*) as well as the collective information state of the community as regards the topic in question may be important. Finally, a search will hopefully lead to a response which constitutes yet another contextual situation (*Outcome*).

The most general work is that of Peter Ingwersen (2007)[53]. His model is abstract and transparent in the way that the above literature-based typology is not. Consequently, it has the important quality of allowing the researcher to recognize types of context not usually thought of, and also is easier to validate. Ingwersen's model consists of a cognitive *actor* (i.e. a user or a team of users) in a **social**, **cultural** and **organisational context**, including tasks and problems to be solved, which interact through an *interface* with a *search system* and a set of *information objects*, together making up the

[53] See also Ingwersen and Järvelin (2005a).

systemic context. Each of these *contextual components* has **intra-object** as well as **inter-object structures**. A document has intra-object structures such as words and phrases, but also inter-object structures such as references to other documents and links to other resources on the same topics. A user has intra-object structures such as relationships between interests in the user's mind and inter-object structures such as relationships to other people. Users also interact with the system during a search session (**session context**) and have an **individual** and a **collective** conceptual and emotional context, of which the former include work task perception and interface interaction whereas the latter covers domain features and other perceptions applied across individuals or across society. Furthermore, there are a set of external **infrastructures** such as features of the network, but also socio-political climate etc. in which the search is couched. Finally, everything has a temporal side to it, the **historic context**, in which developments and changes take place. In order for the full context to be apparent, each contextual component should have all of these seven features analysed (fig. 2.14).

Contextual components	Features
1. Cognitive actors (users)	1. Intra-object features
2. Sociocultural and organisational context	2. Inter-object features
a. Social context	3. Session context
b. Cultural context	4. Individual conceptual and emotional context
c. Organisational context	
3. Systemic context	5. Collective conceptual and emotional context
a. Information objects	
b. Interface	6. Infrastructures
c. System	7. Historic context

Fig. 2.14. Ingwersen's contextual components and features. Each component possesses features of each type which should be evaluated in order to gain a full picture of context.

The important insight behind this model is the fact that contexts can be analysed according to various types of structures that can be applied across all of the contextual components – the centres to which everything else is contextual. This is a really important intuition which should be at the heart of any new attempt at describing context. However, the actual structures need revision and deserve thinking through down to a less abstract level, and so we shall be developing a new context model below along similar lines but with different contents. It is centred on a set of contextual components (but called here **contextual cores**), each with a set of general structures.

Based on the Faceted Stimulus model, and on empirical studies of relevance judgment of contextual features such as the one by Barry (1994)[54], the general typology of contexts illustrated in fig. 2.15 is suggested here. The point of the model developed here is not to suggest actual contextual features that a given system should collect, but to provide a general model of all the *possible* types of context from which a selection should then be made.

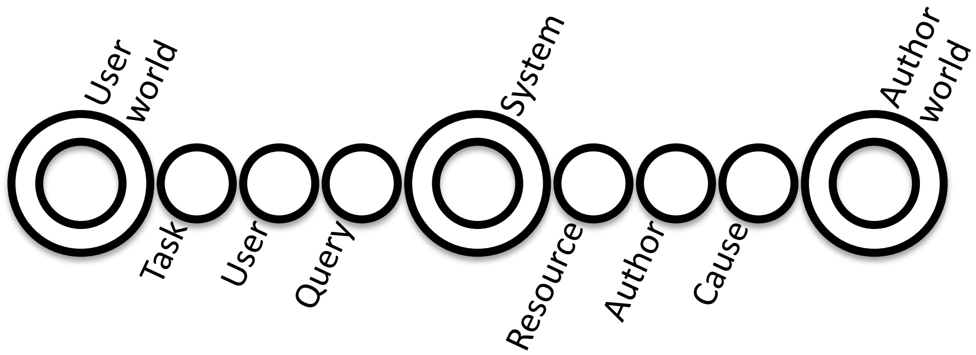


Fig. 2.15. A general typology of contextual spheres.

A LAT needs to be able to interpret the query and the document and then compare them. Any relevant information on the cause and the author are in this context considered part of the document seen as a complex, many-layered entity – the **extended resource** – while relevant information on the task and the user is similarly perceived as part of a complex multi-layered query – the **extended query**[55]. Because the user acts as context to the query, the extended query contains information on the **extended user**, which contains the user including any relevant context. Similarly, as the author acts as context to the resources, the extended resource incorporates the **extended author**, which in turn includes the author with any relevant context. Consequently, there are two main types of context, or *contextual cores*, which we shall call **query context** and **resource context**, respectively.

It is possible to distinguish between different layers of context around the accomplishment of a task, such as the **search** itself, the **investigation** (or **session**) consisting of several searches with the same overall goal, the **work task**, possibly consisting of several investigations, and the wider **socio-cultural context** in which the work is carried out (Ingwersen, 2007; Xie, 2007). However, these are in fact not

[54] See Part II for more information on these empirical studies.
[55] In concrete terms, information on cause and author may be stored as metadata in the document, while information on task and user may be supplied as part of the query or as supplementary information, or inferred from context.

distinct layers but points on a continuum. The only absolute distinction which shall be made here is the one between the interaction with the system, focusing on an immediate aim of getting a certain kind of response, and the interaction with the external world with a particular overall goal in mind, i.e. the **internal** and **external user context**. Some researchers reckon with a separate type of context centred on the device or the system (Ingwersen, 2007; Tamine-Lechani, Boughanem, & Daoud, 2010). In the present analysis we shall consider this the external context in which the resource is couched and/or the search performed, in the same way that users and authors are situated in an external world, which forms part of the context. Thus, it is simply a special set of contextual parameters, which are common to query and resource context. The contextual cores advocated on this background as representing a transparent and illuminating typology of possible types of context are as follows.

1. **Task facets**
    a. **The intentional task**: The task as a moral plight
    b. **The substantial task**: the task confronting the user including the physical user him- or herself, as perceived by the user
    c. **The formal task**: the task as a situation in which the user must find his or her place; something to be analyzed and handled
    d. **The representational task**: the task as being of a certain type and analyzed by the user for causes and effects
    e. **The interpersonal task**: the task as a problem to be solved
    f. **The effective task**: the task as something that affects the user psychologically
2. **Query facets**
    a. **The substantial query**: the query as issued by the user, including any additional information provided to the system
    b. **The formal query**: the query as a linguistic expression
    c. **The representational query**: the query as a semantic description or portrayal
    d. **The interpersonal query**: the query as a message from the user to the system

3. **Resource facets**
    a. **The intentional resource**: the resource as intended for someone
    b. **The substantial resource**: the resource as provided by the author to the system, including any additional information included as metadata
    c. **The formal resource**: the resource as a linguistic expression
    d. **The representational resource**: the resource as a semantic text with portrayals of situations
    e. **The interpersonal resource**: the resource as a message from an author to a reader
    f. **The effective resource**: the resource as something with a psychological effect on the recipient
4. **Cause facets**
    a. **The intentional cause**: the cause as a moral plight
    b. **The substantial cause**: the cause confronting the author, including the physical author him- or herself
    c. **The formal cause**: the cause as something the author must handle
    d. **The representational cause**: the cause as a being of a certain type
    e. **The interpersonal cause**: the cause as a project for the author
    f. **The effective cause**: the cause as something which affects the author emotionally

This makes for 22 distinct contextual cores, each with its own features to note. Each contextual core may be assigned a **structure**, an **environment**, a **history**, and some derivable **consequences**. Structure and environment correspond to Ingwersen's (2007) *intra-* and *inter-object structures*, respectively. In many cases, these are psychological or cognitive structures. Ingwersen's *individual* and *collective conceptual* and *emotional contexts* are here reinterpreted as structure and environment of different contextual dimensions, ordered according to the various contextual cores based on the idea of facets[56], resulting in a far more nuanced and complete description of context. *History* covers the temporal aspect of any contextual core. The *consequences* feature is likely to be the one that is most difficult to understand. It introduces human judgment and cognitive organization from an objective point of view so that contextual data can be ordered into classes such as user types. For instance, the

[56] Note that Ingwersen and Järvelin (2005b) use the term *facet* in a different sense of user vs. system vs. environment, which is here called *contextual cores*.

user's age may not be very interesting in itself, except as a criterion which places the user in the category of child users. As such, this feature may seem to represent an additional operation *on* contextual information rather than the information itself, but then all contextual information is treated in this way when perceived by humans; we invariably judge and organize the incoming stimuli, and this is vital if we are to make sense of them, and this is just as true when the collector of the information is a system.

Perhaps the main advantage of the suggested model is that it concretizes ideas which are present in abstract form in Ingwersen's model so that it can be applied to actual IR systems more easily. For reference, an attempt at a preliminary concretized contextual model with the various features specified is provided in Appendix A. However, further research is needed in order to arrive at a model which acknowledges the different degrees of importance attached to the various types of information and thus supplies a list of information which it pays to collect in a given system.

2.4.4 Relevant context

One obvious difference between search and most other kinds of communication is the mode of expression, which is severely reduced in most queries. Specifically, the way in which the few words of a query should be taken is generally considered exceptionally underdetermined because of the somewhat rudimentary syntax of queries. As we shall see in Part III, there is in fact a surprising amount of syntax in queries: Modification is rife, and the object relation is in frequent use, both with verbs and in nominalized constructions. What is mostly lacking is the subject relation, along with other verb-noun relations normally expressed through case and adpositions. The effect is a kind of unfinished linguistic constructs characterizing the need in some way, but with no concrete explanation of how the expression is to be interpreted in relation to the user or the sought resources. Such relations – which are essential for the system to be able to return relevant information – are left to the powers of inference that current systems really do not have. The upshot of this is that, as is true of linguistic messages in general, queries do not code all the necessary information. In fact, a large proportion of their meaning has to be extracted from context. A query only provides *evidence of the user's intentions* to convey a certain meaning, while the rest has to be inferred from context. The traditional IR approach implicitly analyses a query as a simple code which incorporates the entire message, and this is why it can never decode its true meaning; it is essential to gather supplementary information from users (Swanson, 1977). The query is not merely a string; the actual expression is only the tip of the iceberg. A full query should not only include all of its facets, but also all the contextual information that the addressee is expected to draw on and exploit for making inferences. The question is of course where to stop; if the query is merely a clue then interpreting what the user really wants becomes highly subjective. In other words, it becomes a question of judging what is *relevant*. Part II is an exploration into this still somewhat uncharted territory.

Saracevic (1975) seems to have been the first to state explicitly that search is an instance of communication and to develop a theory around this insight. Unfortunately, as he himself pointed out later, his model was built on the simple code model of communication which was the only one around at the time, and thus it "implies, but does not incorporate the inferential, interactive communication exchanges. In other words, while the framework is still good for an inventory of elements and relations involved, it is woefully inadequate for explication of the interactive dynamics of relevance in IR processes" (Saracevic, 1996, p. 7). With the development of Relevance Theory, a new possibility of forming an adequate theory of relevance in IR has presented itself.

Relevance Theory developed as a reaction to Grice's inferential model. Its proponents – starting with Sperber and Wilson (1986) – did not feel that cooperation was a sufficiently strong agent to explain relevance in all its shades. Instead, they claimed that a *cognitive principle of relevance* has guided human development and shaped our consciousness in a way which prioritizes relevance (see chapter 4).

The cognitive principle of relevance: Human cognition tends to be geared to the maximisation of relevance

Relevance Theory claims that the way we interpret messages is by testing the suitability of different interpretations in order of accessibility, i.e. the most obvious and likely interpretation is assumed first, and then alternatives are considered in turn. We stop this process when our expectations of relevance are met, i.e. when we find a way of interpreting the message that seems likely, and which yields a positive cognitive effect. A speaker who wants her utterance to be as easy as possible to understand should thus formulate it so that the first interpretation to satisfy the hearer's expectation of relevance is the one she intended to convey. With this in mind, it is clear that there is another difference between straight-forward question-answer communication and the results list as an answer to a search request. The results list is not in itself an instance of communication from an intending speaker; it is an automatically compiled list of messages produced by different "speakers". In communication, there is rarely more than one response to any one utterance; if a question is asked, usually only one answer will be given at a time, so no choice has to be made in selecting one to attend to. However, results lists break the principle of not producing more than one relevant input simultaneously. If we are to take

Wilson and Sperber's claim literally, search results which are not at the top are less relevant due to the very fact that they are presumably slightly less accessible to the user than the one at the top and hence require more efforts to get to. If this is true then researchers who evaluate result ranking are facing a circularity problem, since what is shown at the top automatically *becomes* more relevant than the rest, simply by being presented first. Note, also, that it is not necessarily the case that what is presented at the bottom is less accessible than what is presented immediately above it, since there might conceivably be a peripheral effect so that what is at the bottom is slightly more attention-grabbing than what is right above it, depending on the screen layout. Consequently, a results list may generate its own field of relevance.

It is suggested here that the cognitive principle of relevance is so deeply rooted in our instincts that even though users of a LAT know that they do not submit their queries to a human addressee, they cannot help observing the cognitive principle of relevance, which keeps them from providing what would otherwise have been excessive information inferrable from context. Consequently, LATs are compelled to do the inferring as if they were human interpreters. In other words, we might as well attempt to design LATs to work as human listeners, since this is the kind of communication partner that people understand best and are best at supplying with the necessary information.

> First and foremost, system designers should look into designing information retrieval systems that afford the opportunity for searchers and systems designers/indexers to converse in a more real-time mode. (Blair, 2003, p. 39)

This is not to say that users must necessarily provide whole natural-language sentences for such a system to work; it is more a question of making systems able to understand what people try to communicate, and if this requires that users provide some basic information about their needs apart from the usual queries, then such information ought to be collected[57]. Here is an example from one of the analysed logs.

2.10	**eastern europe**	WID (19)

This query was issued at the Workindenmark job portal. These users were not interested in information on Eastern Europe, but rather wanted to inform the system that they were from Eastern Europe, or possibly that they wanted to work in that part of the world for a Danish company. In this particular

[57] The empirical study by Kelly, Dollu and Fu (2005) provides suggestive evidence that people may indeed be more willing and/or able to communicate their needs in natural language than in the form of keywords, even when asked to do so in writing. However, they caution that this particular result may possibly be due to other factors.

case, the search is likely to be successful to the extent that such a job is available. This is because the domain context is so narrowly restricted in the WID site that there are not likely to be many irrelevant resources on the topic anyway. Most search engines provide access to resources belonging to wider domains than this, and then the unwillingness to provide necessary contextual clues can be highly problematic[58]. Question-answering (QA) systems encourage users to employ full sentences, which they interpret with varying degrees of linguistic accuracy. Most systems, however, are designed to accept one or a few syntactically unconnected keywords. Surprisingly often, users do provide a traditional search engine with full questions or sentences (see section 9.3.2). This frequently results in a much better description of the user's information needs and would be preferable to the extent that the LAT were able to interpret this input. Some kind of fusion between search engines and QA systems might provide the system with the information necessary to perform a helpful search.

Having said this, it should be added that there is no guarantee that aiming to produce a LAT which would interpret the query exactly as a human would is a good idea, even if it were possible, because queries are not expressed exactly as they would be if the addressee were human and consequently there is no comparable process in natural language that might be studied in order that we might understand the process. Thus, rather than studying context from the point of view of interpretation, which might intuitively be thought the most straight-forward thing to do, IR research will presumably profit more from an exploration of the production aspects of context. This way, the contextual parameters that cause users to express themselves in a certain way can be studied so that the reasoning behind the expression may be understood. If the contextual information that is crucial for the production of a query can be identified, it can be reasonably supposed that this same information should be necessary for the interpretation of it, though difficulties in connection with interpretation may well differ from those of production. This is the approach that we shall be following in Part III.

2.4.5 Introducing an intermediary function

In the library research tradition, it is normal to consider the information retrieval process to involve an **intermediary** – originally a librarian, but potentially a computerized replacement (Saracevic, Kantor, Chamis, & Trivison, 1988). The role of this intermediary is to help clarify the focal problem and formulate a request that is likely to yield the desired kind of response from the system. Unlike the user, the intermediary is supposed to know the system and to be an experienced and proficient user of the

[58] It is likely, of course, that this particular request would have been expressed using a slightly longer query in a web search environment. However, even though "it is obvious" what the user's intention is in a narrow domain such as Workindenmark, it is in fact not at all obvious to the system, and the user has to rely on the absence of irrelevant material relating to other intentions.

tool. In mainstream IR research, this intermediary is dispensed with, seeing as most searches are performed without a human intermediary these days. However, it may be a good idea to reintroduce the intermediary in the conception of future search systems if we are to improve them (Ingwersen, 1992).

In a sense there is always an intermediary, although the role can be taken over to some extent by the user and/or the system, depending on the sophistication of the latter. Thus, I will suggest that the **requesting user** that we defined above under the name of *user* as the association of a request, a search and a need, should be kept distinct from the **intermediating user**, who interacts with the tool. The interaction is both psychological and concrete in that it consists in 1) clarifying the problem to be solved (as experienced by the requesting user), 2) deciding on a compromised need after considering the capabilities of the tool, 3) formulating a query, 4) delivering the message, 5) receiving the response, 6) evaluating the relevance of suggested resources, and 7) delivering the result to the requesting user. The latter is the one who 1) experiences the problem, and 2) applies the result. Prototypically, of course, the requesting and the intermediating user reside in the same body, but they are conceptually distinct; one is an experiencing psychological individual and the other is an acting body.

The intermediating user is notably also distinct from the **intermediary function**, because the latter is defined as the agent who deals with the translation of the visceral need into something that will trigger a useful response. This is a *function* which can be distributed across the intermediating user and the system. Information retrieval involves various distinct functions, most notably a **requester function** held by the person who experiences the need, an **intermediary function**, which can be split between the same person or a different person on the one hand and the system on the other, along with several others, such as the **matching function** held by the system, the **indexing function** held either by a system or a person, an **access function** held by the system and a **relevance evaluation function** typically split between the intermediating user and the system.

It is often more appropriate to speak of *functions* rather than *users*. However, the word *user* is so common and springs so naturally to mind that we shall not desist from using it unless we explicitly wish to draw attention to the various functions.

3 Optimizing search efficiency

The success or failure of any interactive system and technology is contingent on the extent to which user issues, the human factors, are addressed right from the beginning to the very end, right from theory, conceptualization, and design process to development, evaluation, and to provision of services.

(Saracevic, 1996)

Apart from exploring the search process, it is vital for our project of understanding users' reasons for expressing themselves in certain ways that we find a way of studying user behaviour in the first place. Introspection can only bring us so far, since it is likely that different users do not express themselves, or even experience the search process, in similar ways. Consequently, it is necessary to access statistical data in some way, and we shall choose to do so through search logs obtained mainly from two sites powered by Ankiro Enterprise Search. In the ideal case, such data should be supplemented by some form of interviews where users could give their own explanations of their behaviour, to the extent that they are able to do so. However, that would require a focus on a much smaller dataset, leading to less reliable statistics overall. Instead, we shall rely on the transaction logs to provide the necessary data while remaining true to the surface expression. We shall attempt to avoid assuming any interpretations that are not explicitly expressed in the queries under analysis and simply consider ambiguous queries to be just that, rather than assign them to the most likely type, but much more will be said about this in Part III. Section 3.1 introduces the two search logs selected for the main empirical analyses in this thesis. They were chosen for their substantial differences in order to achieve a maximally broad coverage. Naturally, this brings its own problems of incompatibility of results between the logs. However, this may be a very positive side-effect, since it reminds us of the importance of not generalizing results too much across domains.

Even if users do reach a reasonably full understanding of their needs, the task of actually coming up with an effective query still remains. This involves on the one hand reflecting the need suitably well and on the other anticipating what form the query should have in order to produce the best response possible. Based on the log material as well as results published previously by other researchers, section 3.2 looks at how many words users employ when describing their needs in a query, and what the maximally efficient query would be like. It turns out that even though queries are typically very short and ambiguous, users are in fact quite good at providing efficient queries. Interestingly, the data also show that user behaviour in this respect varies with the type of query. While query length is clearly a very indirect indicator of user intentions, this does suggest that we need to look further into the distribution of various query types, which is the purpose of Part III.

The discovery that search is a process that the user must go through in order to find out what he or she really needs should have consequences for the design of information tools. Over the years, various researchers have suggested ways of adapting search technology to the human way of interacting with IR tools, and section 3.3 discusses a few of them.

3.1 Studying user behaviour through query logs

One way of studying search behaviour is by analysing **query logs** (also known as *transaction logs*). A log file typically stores all queries issued along with the time of issue and the IP address of the user. This helps the analyst recognize sessions of interrelated queries representing one user struggling to find the necessary information. Frequently, click behaviour information is also included, showing what resources the user chose to consult after being presented with a list of results, and sometimes even the time spent on viewing the documents. This may help identify the documents that the user thought likely to provide relevant information, and also to some extent whether this assumption was justified – if a page was only viewed for a few seconds it is likely that it was immediately obvious to the user that it would not provide useful material and hence that different resources should be preferred in response to that particular query in future.

3.1.1 Data

Log material for the main analyses in this thesis was drawn from two different sources. This was done in order to achieve a broader coverage than one log can offer. The first source used is the Danish homepage of a private enterprise of sufficient size to provide a sizeable log, consisting of 2,614,701 searches performed during one year (2010-2011) on the homepage of TDC, a major Scandinavian telecommunications company, and henceforth abbreviated to L^{TDC}. Queries were principally issued in Danish. Services include product sale, support with problems, and corporate and other information from the site. The kind of queries issued is thus relatively diverse compared to both a strict e-shop and a support or information site.

The other material, L^{WID}, is derived from Workindenmark, the site of the Danish Agency for Labour Retention and International Recruitment, and consists of 204,546 searches performed during the year 2011, mostly in English but also partly in Danish and various other languages. Workindenmark is a public service provided in English to people of other nationalities looking for employment in Denmark. On this site, searches are accordingly typically oriented towards solving the particular problem of getting a job. The problems and solutions are much more personal than in the support situation concerning electronics, and it was interesting to see that this is reflected in the way people use language in expressing their needs. Furthermore, the simple fact that the domain is so much narrower affects query expression, because it means that contextual information can be left out to a much greater extent. The choice of these two sites provides a contrast in the main language used (Danish vs. English), in the private vs. public dimension, and also in the domain of content.

● The TDC log

The full TDC search log analysed contains 2,614,701 searches distributed across 377,384 distinct[59] query strings issued during one year. The queries are almost exclusively in Danish. Unfortunately, I did not have access to information on paging, and consequently several consecutive identical searches may represent the same user viewing several pages. As the effect is evenly distributed across the dataset, however, it is not likely to affect any of the findings; it simply means that some numbers may be slightly exaggerated. Before calculating the above figures, a total of 127,353 searches for the string *mobil* 'mobile (phone)' from only two IP addresses were removed, since they most likely represent monitoring robots. Even after removing the robot-generated queries, the query string *mobil* still represents 9% of the entire dataset, or 239,529 searches. Together, the 10 most frequently occurring strings account for 22% of the search log and the first 100 represent 41% (fig. 3.1). On average, each string occurs 6.93 times, but 70.8% occurs only once[60].

[59] The term *distinct query* is used here of the collection of all searches where the same query string has been issued. Some authors appear to use *unique query* in this same sense while for others it appears to have the opposite sense. Consequently, that term is avoided here.
[60] This figure includes all remaining queries without further filtering.

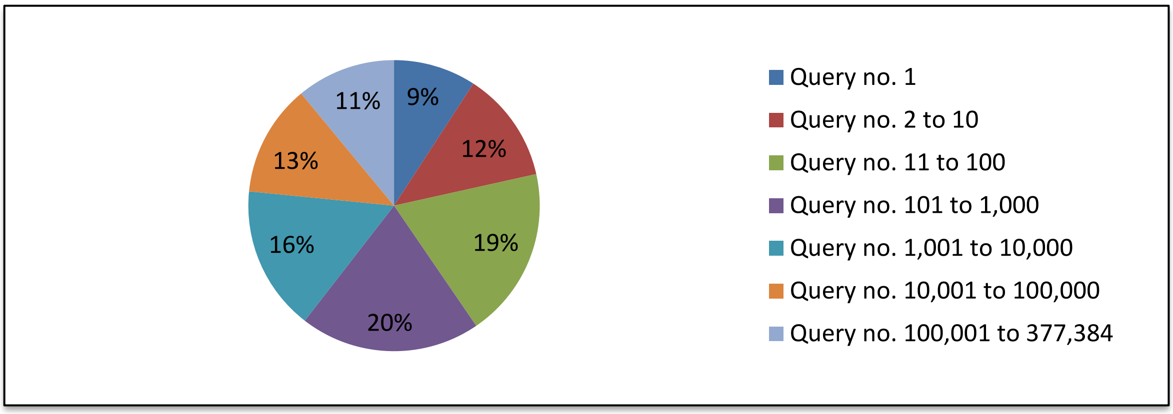


Fig. 3.1. The distribution of distinct query strings in L^TDC^. The most frequent word, *mobil*, represented 9% of all queries issued during one year, even after cleaning away a quarter of a million instances which seemed to be due to monitoring robots. The rest were distributed in a roughly logarithmic manner.

● The WID log

The full WID log file contains 204,506 searches distributed across 50,449 distinct queries issued across one year. Most queries are in English, but a few are in various other languages (fig. 3.2). There are no obvious signs of robot activity. The distribution of queries in this case differs markedly from L^TDC^, with no prominent peak in the most frequently used words, and consequently no logarithmic distribution, at least among the 100 most frequent queries (fig. 3.3).

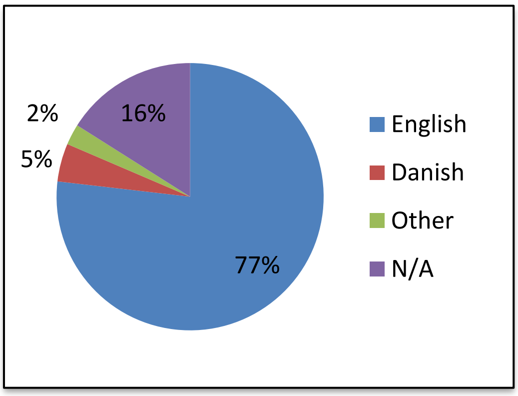


Fig. 3.2. The distribution of languages used in searches performed on the WID site in 2011. 16% of all queries only contain names of places, people, tools, methods, etc. which do not represent any language.

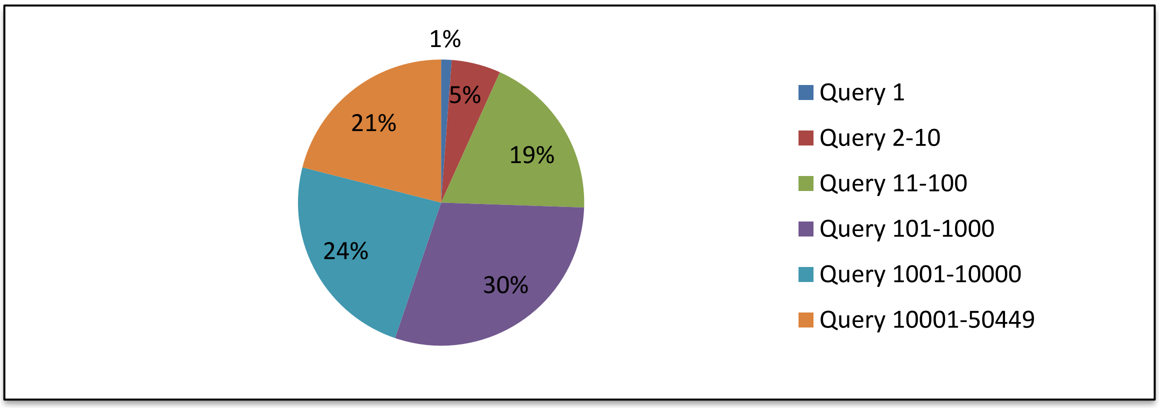


Fig. 3.3. The distribution of distinct query strings in the WID log material.

> On average, each string occurs 4.05 times, and 75.1% of them just once. Clearly, there is more diversity in this log file than in L^{TDC}. In this dataset, careful attention was paid to paging, allowing it to be used for the Target Analysis in chapter 10.

3.1.2 Reducing the size of a search log

The considerable size of the search logs used meant that analysing all queries manually would have been entirely unrealistic. Consequently, for the manual analyses (the Target Analysis in chapter 10 in particular) the WID log was reduced to 10% using a specially developed method.

The search logs used were organized in such a way that all identical queries were represented only once, with a note of its original frequency. This a very common way of dealing with log data which has the unfortunate consequence that extracting part of the log in an unbiased way becomes somewhat cumbersome. The method devised for this extraction process runs as follows.

> In order to reduce a log to *p*% of its original size, all queries that occur at least 100/*p* times have their logged frequency divided by 100/*p*. Consequently, they only count *p*% as much as in the original dataset. From the remaining queries, only a portion of the queries is included in the analysis, while the rest are discarded. At each level of frequency, $(p/100)*f*n$ randomly selected queries are included, where *f* is the frequency and *n* the number of distinct queries at that level. For instance, in a reduction to 10%, only 50% of the queries of frequency 5 are included in the analysis. All of these are assigned a frequency of 1, regardless of their true frequency.

The result is that while each less-frequent query is assigned too much weight, every frequency level has the correct relative weight. Since search behaviour varies significantly between the most and the least frequent query types, this is important; no analysis that only includes the most frequently occurring queries can be deemed adequate because it may well turn out that the least frequent segment together represents one extremely important category that simply shows a large amount of variability in expression. Phone numbers constitute such a category in the TDC log; many users search for phone numbers, but rarely for the same ones.

The WID log was reduced to 9277 distinct queries corresponding to 10% of the searches. The number of queries to be analysed was hence reduced to about 19% of the original number with this method. It should be noted that reducing a query log in this way naturally precludes any session-based analysis, as sessions are freely broken up, and consequently no such analyses are attempted in this work.

3.1.3 Query types

The most basic type of user behaviour that is easily measured is the number of words in a query. It is frequently remarked that queries tend to be short and ambiguous, and in fact we shall see that the statistics turn out to be remarkably stable across experiments, something which has not to my knowledge been recognized before.

It would be interesting to compare the results arrived at in connection with enterprise search to web search, and also to compare old data to new results. Unfortunately, it is not possible to isolate one effect from the other, and there are several other parameters that differ and influence the results, so that it is in fact difficult to compare the various figures. The results from L^{TDC} and L^{WID} may of course be compared, but even here there is a difference in main language as well as domain. Even so, the figures in themselves are sufficiently interesting to warrant the results of the statistical study to be presented here. Before this can be done, however, the datasets need to be cleaned.

Search logs are notorious for being “messy” (Jansen, Spink, Bateman, & Saracevic, 1998), i.e. they contain a lot of uninterpretable queries as well as a huge amount of spelling errors and aberrant language use, in addition to being “noisy” (Grimes, Tang, & Russell, 2007), i.e. including robot-generated queries, spam queries, hacking attempts, etc. This makes linguistic analysis difficult, because such analyses must be performed automatically due to the immense number of queries studied. This is another reason for filtering the search strings according to their nature in order that clearly non-linguistic queries could be removed. Queries discarded without further ado as being irrelevant to the study amounted to little more than 5% of searches in L^{TDC} but represented more than 16% of the distinct strings submitted.

Counting the words in all queries does not really make sense, because some queries do not actually contain words at all. How is a phone number to be counted, for instance? Is it one word or several if it has been written with spaces in between the numbers? This is especially tricky as the same phone number may be written with or without the spaces. Since L^{TDC} contained more than 90,000 queries that only consisted of a phone number, the problem simply could not be ignored. Consequently, the queries were grouped into a range of types to be counted separately, as follows.

- Clearing the search box failed
- Phone numbers
- IP addresses
- Other numeric queries

- E-mail addresses
- References to the homepage itself or a subpage
- External URLs
- Incomplete URLs
- Long queries (10 words or more)
- Normal queries (the residual set)

A computer program was developed which was able to analyse large amounts of queries according to these criteria. The individual types are explained below, along with statistics for L^{TDC}.

Clearing the search box failed

In L^{TDC} 29,227 searches (1.11%) contained a string advising the user to search the site in question ("Søg på TDC.dk"), often with other search terms inserted somewhere in the middle (25,982 of the searches or 88.67% of them did not contain a query at all). These are caused by the fact that the search box contains this string by default, and sometimes it is not cleared before a query is entered. Since such searches are likely to be repeated as soon as the user realizes what has happened, they were excluded from the dataset.

Phone numbers

90,191 searches (3.44%) were (or looked like) phone numbers. The phone number for directory service *118* was especially common at 15,775 searches (17.22% of the phone number searches)[61]. Phone numbers were excluded from the calculations of query length.

IP addresses

536 searches (0.02%) seemed to be IP addresses. These were likewise excluded from analysis.

Other numeric queries

3,468 searches (0.13%) were numeric without belonging to any of the above categories. These, too, were excluded.

E-mail addresses

4,907 searches (0.18%) were e-mail addresses. These were excluded from analysis since they are likely to be due to a misunderstanding of what the system is capable of. This may of course not be true in every case, but it was nevertheless decided to skip them. All queries

[61] It is possible that this was in many cases not considered a number but more of a name of the service.

containing a “@” were indiscriminately registered by the system as e-mail addresses, which may in a few cases have been wrong, but the number of errors has been found to be vanishingly small.

References to the homepage itself or a subpage

9,003 searches (0.34%) were partial or entire URLs referring to the homepage itself or its subpages. These were excluded from analysis since they are likely to be due to a misunderstanding of what the system is capable of. They were identified through their containing the name of the homepage in combination with the string “.dk”. There are a few cases where these represent normal queries, but the number of errors is small.

External URLs

2,499 searches (0.09%) were URLs referring to external web locations. These were excluded from analysis since they are likely to be due to a misunderstanding of what the system is capable of and/or the user has mistaken the search box on the site for a web search box. External URLs were identified by looking for the string “www.” in the residual set after removing the above-mentioned types. These were all correctly identified.

Incomplete URLs

8,044 searches (0.30%) were partial URLs referring to external domains. These were excluded from analysis since they are likely to be due to a misunderstanding of what the system is capable of and/or the user has mistaken the search box on the site for a web search box or URL box. The difference from the previous categories is that these are identified as queries containing “.dk” or “.com” without either the name of the homepage or the string “www.”.

Long queries

1,004 searches (0.04%) corresponding to 800 distinct strings (0.21%) were “normal” queries but contained 10 words or more. Since most queries with 10 words or more were not really queries in the true sense of the word but rather sentences that had been entered due to a misunderstanding of what the system was capable of, these might reasonably be excluded from calculations and further analyses. Many such “queries” were clearly intended to be read and answered by a human recipient.

Residual normal queries

The remaining 315,881 distinct strings (83.70%) representing 2,465,822 separate searches (94.31%) were normal queries in the sense that they did not belong to any of the above categories. These were the ones chosen for the L^{TDC} analyses in this thesis.

The statistics for the various types can be found in fig. 3.4.

	No. of searches		Distinct strings	
Clearing the search box failed	29,227	1.11%	2,100	0.55%
Phone numbers	90,191	3.44%	48,193	12.77%
IP addresses	536	0.02%	141	0.03%
Other numeric queries	3,468	0.13%	1,296	0.34%
E-mail address	4,907	0.18%	3,172	0.84%
Homepage or a subpage	9,003	0.34%	1,988	0.52%
External URLs	2,499	0.09%	1,213	0.32%
Incomplete URLs	8,044	0.30%	2,600	0.68%
Long queries	1,004	0.04%	800	0.21%
Residual normal queries	2,465,822	94.31%	315,881	83.70%
Total	**2,614,701**		**377,384**	

Fig. 3.4. The distribution of different query types in L^{TDC}. The percentage of distinct queries has been included for completeness, though the number of actual searches is far more interesting.

The analysis of L^{WID} was simpler, as the proportion of non-linguistic queries was much smaller in this log. In the L^{WID} analysis, the numeric category was made to include phone numbers, postal codes and IP addresses as well as other numeric queries, as each was quite small. Numeric searches amounted to 1,396 searches (0.7%) in total. A special category of queries in other scripts than Roman was added, and these were removed from the technical analysis. The full statistics can be found in fig. 3.5.

	No. of searches		Distinct strings	
Numeric queries	1,396	0.68%	924	1.83%
E-mail address	167	0.08%	132	0.26%
Homepage or a subpage	0		0	
External URLs	679	0.33%	465	0.92%
Incomplete URLs	92	0.05%	79	0.16%
Long queries	0		0	
Other scripts	151	0.07%	151	0.30%
Residual normal queries	202,061	98.8%	48,698	96.5%
Total	**204,546**		**50,449**	

Fig. 3.5. The distribution of different query types in L^{WID}.

L^{WID} can be seen to be much cleaner than L^{TDC}, which is not surprising given that the site has less diverse uses and fewer users altogether.

3.2 The efficient query

The results of the basic query length analysis can be read from the tables in fig. 3.6 and 3.7.

No. of words	No. of searches		Distinct strings	
1 word	1.693.891	64,78%	127.166	33,70%
2 words	589.404	22,54%	111.824	29,63%
3 words	238.315	9,11%	78.632	20,84%
4 words	59.373	2,27%	35.095	9,30%
5 words	19.986	0,76%	13.943	3,69%
6 words	7.144	0,27%	5.474	1,45%
7 words	3.140	0,12%	2.506	0,66%
8 words	1.521	0,06%	1.198	0,32%
9 words	755	0,03%	616	0,16%
10 words	429	0,02%	331	0,09%
>10 words	730	0,03%	591	0,16%

Fig. 3.6. Distribution of queries of various lengths in L^{TDC} as measured by the number of words they contain.

No. of words	No. of searches		Distinct strings	
1 word	145,862	71.31%	21,632	42.87%
2 words	46,969	22.96%	19,481	38.61%
3 words	7,969	3.89%	6,095	12.08%
4 words	2,077	1.01%	1,816	3.59%
5 words	756	0.36%	627	1.24%
6 words	298	0.14%	279	0.55%
7 words	187	0.09%	173	0.34%
8 words	82	0.04%	76	0.15%
9 words	77	0.03%	69	0.13%
10 words	69	0.03%	60	0.11%
>10 words	200	0.09%	141	0.27%

Fig. 3.7. Distribution of queries of various lengths in L^{WID} as measured by the number of words they contain.

It is immediately clear from these tables that short queries are strongly favoured by users of the sites analysed here. In fact, one-word queries outnumber even two-word ones by far. However, this depends on whether all searches or only distinct queries are counted, the reason being that longer queries are significantly more likely to be distinct in the log (fig. 3.8-3.11).

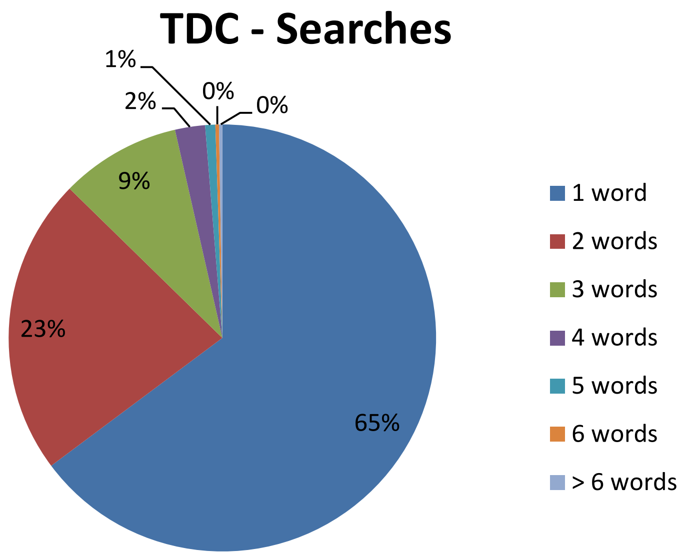


Fig. 3.8. Distribution of separate queries of various lengths in L^{TDC} as measured by the number of words they contain.

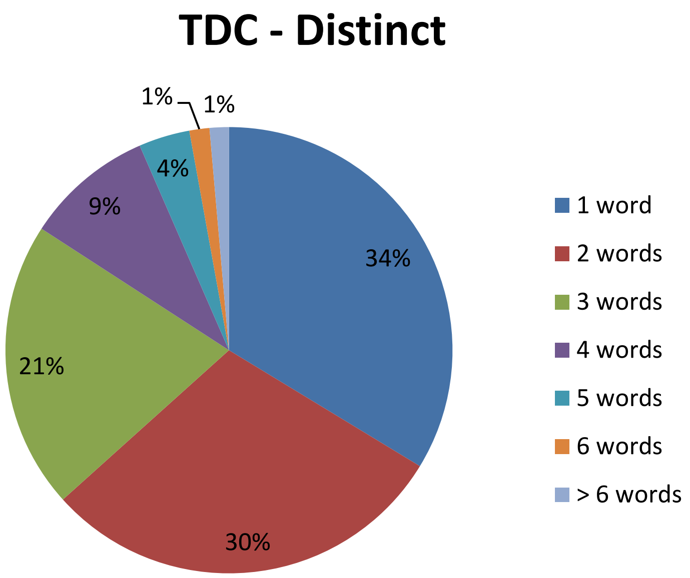


Fig. 3.9. Distribution of distinct query strings of various lengths in L^{TDC} as measured by the number of words they contain.

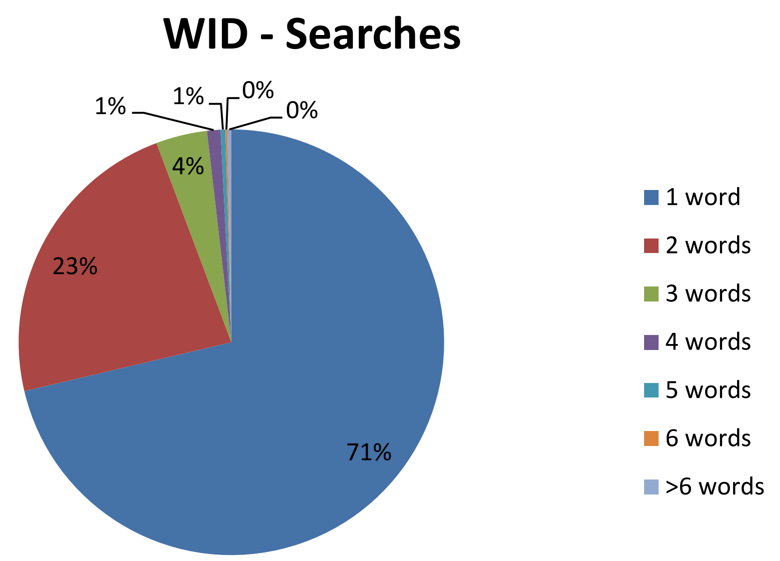


Fig. 3.10. Distribution of separate queries of various lengths in L^{WID} as measured by the number of words they contain.

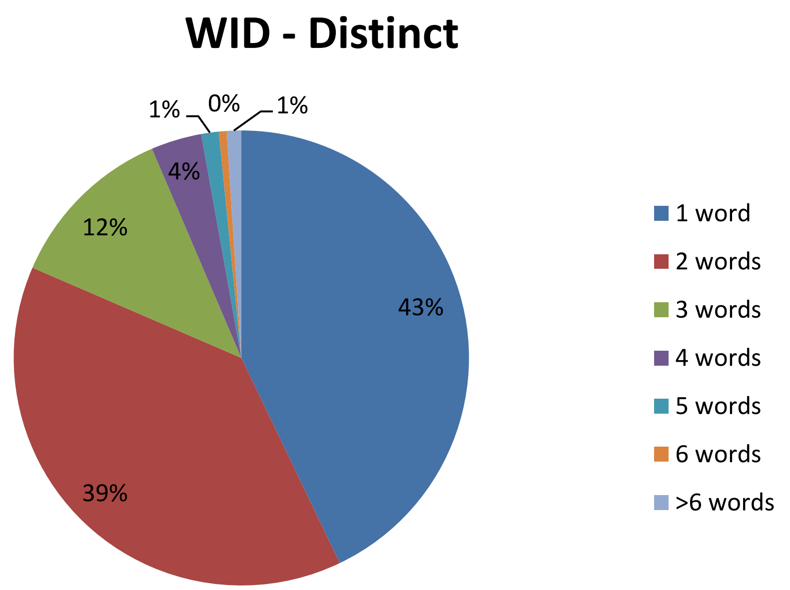


Fig. 3.11. Distribution of distinct query strings of various lengths in L^{WID} as measured by the number of words they contain.

The average number of words in a query was calculated to be 1.54 in L^{TDC} (or 1.53 if all query types were included). Excluding or including queries marked as long did not change this figure. If only distinct queries were counted, the average rose to 2.30 (or 2.51 with all types included). This is because short queries are overrepresented among the frequently recurring queries.

In L^{WID}, the average was 1.37 (1.38 with all queries included), which is even shorter than usually reported. This result is not surprising given that the domain is quite narrow so that contextual information can be left out when formulating a query. For instance, a user looking for a job as a strawberry picker may simply enter the query *strawberry*, which in a broader domain would be unlikely to be precise enough to result in any job postings.

3.1	**strawberry**	WID (64)
3.2	**strawberry picking**	WID (10)
3.3	**strawberry picker**	WID (6)
3.4	**strawberry pickers**	WID (2)

However, as we shall see below this result seems to stand out as exceptional among search engines of many varied types.

3.2.1 The Standard Query Length Distribution

It is possible to count either all separate searches performed during the period investigated, or to count only the distinct queries. Thus, the fact that 15,775 people searched for *118* in L^{TDC} (or rather, *118* was searched for at this number of separate occasions, to be exact) could be counted either as one instance of a phone number, or alternatively as 15,775 such instances. The latter option is the more informative in most circumstances, since it involves the number of times this particular query was in fact issued. However, some kinds of analysis do require the counting of distinct queries. This is because even though every single phone number is not very significant (disregarding the exceptional case of *118*), phone numbers as a class are of great importance to users of this site. Thus, users are served best if the search engine is able to handle phone number queries. If all phone number searches are counted separately, this considerable need is not apparent. In fact, only 4% of searches are phone numbers (excluding the large number of searches that combine phone numbers with the names of people or their addresses). This is of course not a trivial amount in itself, but if the calculation is based on *different* phone numbers, the percentage rises to 13% of all distinct query strings issued (fig. 3.12-3.13). This segment can hardly be disregarded as insignificant. The distinction between searches and distinct queries should be borne in mind when existing studies are consulted.

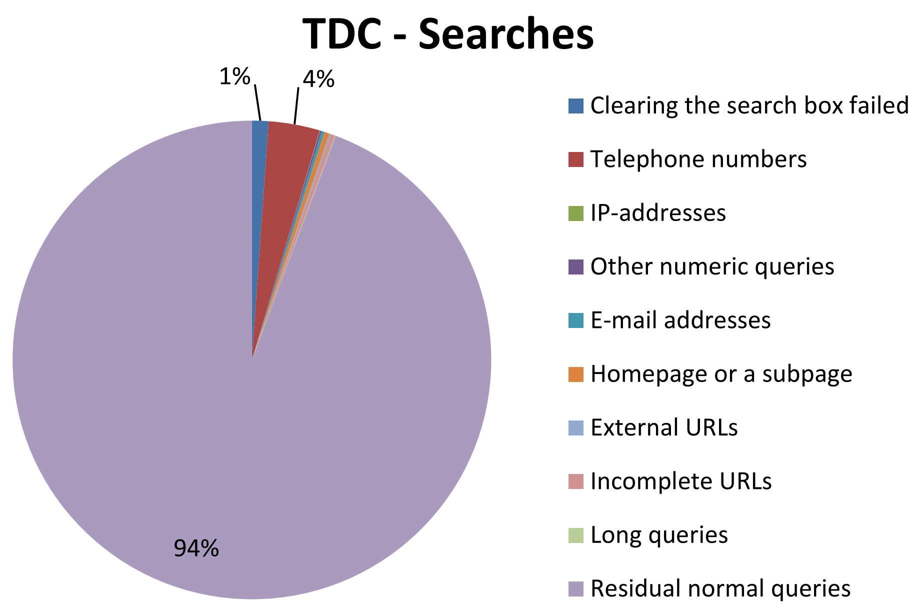


Fig. 3.12. The distribution of different query types in L^{TDC}.

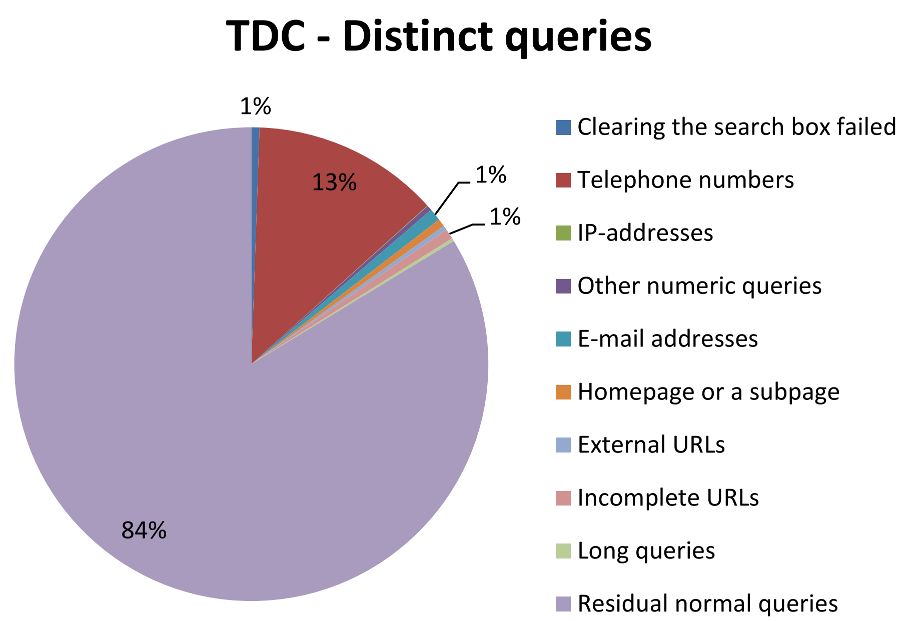


Fig. 3.13. The distribution of different query types in L^{TDC}.

In previous work, the average query length in web search has been found to be around two to three words, and very few queries have more than 6 words (Jansen, Spink, Bateman, & Saracevic, 1998; Zien, Meyer, & Tomlin, 2001). As can be seen from fig. 3.14, most tend to contain only one or two words (Jansen & Pooch, 2001)

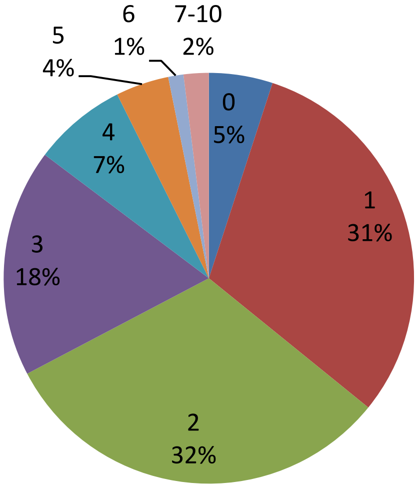


Fig. 3.14. The distribution of queries by the number of terms they contain in the sample investigated by Jansen *et al.* (1998). The study involved 51,473 queries from the Excite web search engine. Queries of length 0 indicate that the user has exploited the search engine's "More like this" feature. The diagram is based on data provided in the paper.

An extensive study performed by Silverstein *et al.* (1999) on more than 575 million queries found a distribution that – when the special 0-length queries explained in fig. 14 are disregarded – is almost identical to the one reported in Jansen *et al.*, despite the much larger dataset (fig. 3.15).

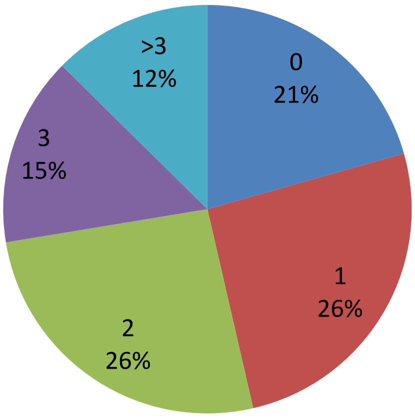


Fig. 3.15. The distribution of queries by the number of terms they contain in the sample investigated by Silverstein *et al.* 1999.

Interestingly, the distribution reported in these studies is furthermore almost identical to the one arrived at for L^{TDC} (fig. 3.16). It can thus be taken to be fairly well documented that this is indeed the distribution of query lengths in general in web search, and perhaps even in enterprise search. It would seem that one- and two-word queries as a rule cover one third each, while three-word queries account for about 20% and the remaining 14% are longer. The consistency of these results is highly remarkable.

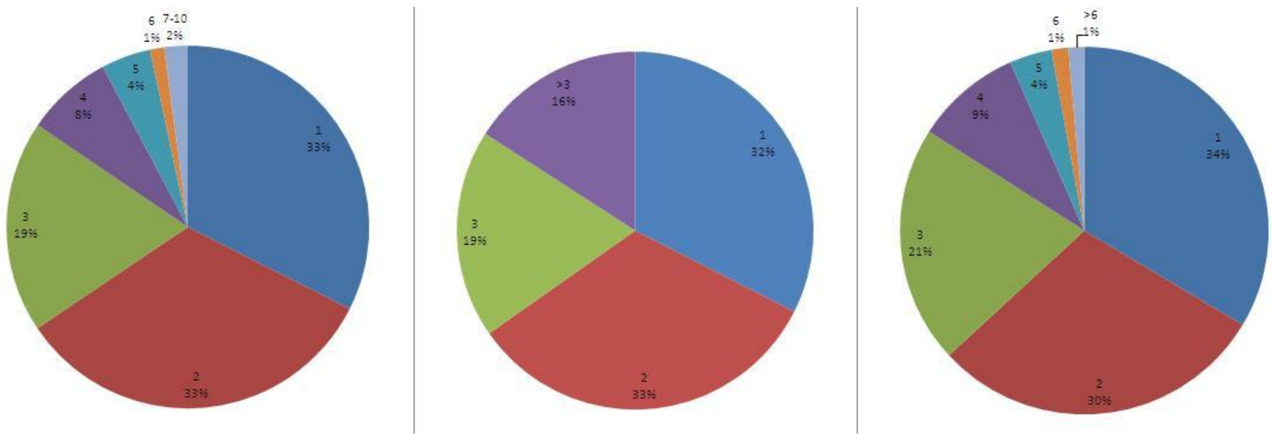


Fig. 3.16. A comparison of the query length ditributions of distinct queries found in Jansen *et al.* 1998, Silverstein *et al.* 1999, and the present study of L^{TDC}, respectively. Silverstein *et al.* do not distinguish between the various query lengths above 3. In departure from the diagrams shown in the papers, all empty queries have here been removed from the statistics for comparison of actual queries only. The result is an almost identical distribution in all three studies.

L^{WID} deviates from this distribution with 82% of queries containing at most two words; the reason is most likely to be sought in the much narrower domain of this site, perhaps augmented by the lower levels of confidence among the international users searching in English, which is clearly often not their native language.

Silverstein *et al.* report an average word count of 2.35, while Jansen *et al.* obtain an average of either 2.35 or 2.21. It is unclear what the difference between these two numbers is, but it may be a difference between distinct and non-distinct queries, or one of them may simply be a mistake. If the figure of 2.21 is comparable to our 1.53, then there is a substantial difference, in spite of the fact that the search engine Excite studied by Jansen *et al.* did not allow queries longer than 10 words, whereas the one under study here allows indefinitely long queries, with a few queries having as much as 50 words (though these hardly qualify as true queries). On the other hand, 2.35 is strikingly similar to our 2.3/2.51 (depending on the amount of filtering).

Silverstein *et al.* indicate explicitly that they counted distinct queries only, though why they do this is not at all clear. Counting the distribution of query lengths across all searches would seem to be so much more interesting. Jansen *et al.* do not explicitly indicate whether they counted distinct queries only or all searches, but their results seem to indicate the former, since the identical outcomes in fig. 16 can hardly be coincidental. If this is so, then it is highly remarkable that the results arrived at in 1998 with an English-language web search engine should have been replicated 13 years later with a Danish-language enterprise search engine. This is hardly expected. If, on the other hand, Jansen *et al.*'s figures do in fact refer to all searches and not just distinct strings, then the result is indeed highly dissimilar between the two analyses.

It should be noted that – quite apart from the question of what the Jansen *et al.* study really implied – the present analysis is not immediately comparable to the data used by the other studies, for several reasons. Firstly, their studies were based on web search, while this one was directed towards enterprise search. This difference would in itself be highly interesting to study. However, the other differences make such comparison difficult. One additional difference is that many queries in the logs studied here were in Danish, which is a problem because Scandinavian languages have different compounding strategies from English. Compounds are mostly written solid in Danish, which reduces the number of words significantly. In other words, a difference in mean query length could simply reflect a linguistic difference rather than one between web and enterprise search. There is a relatively large amount of inconsistency in L^{TDC} as regards compounds, i.e. the same compound is frequently found written both as one and two words. Thus, if users had followed Danish grammar consistently, the average length would be even shorter.

In order to see the effects of language on query length, we turn to a third query log which contains search data from several countries and thus presumably in different languages. The log that we shall be using is that of the search engine employed internally in the Denmark-based international enterprise Novozymes. These L^{NZ} transaction logs contain information on the country in which a given search was performed. This means that we can compare the queries originating in Denmark (which make up 64 % of the log) to the remainder, which is largely in English, with a few interspersed queries in other languages. Unfortunately, a lot of search is performed in English in Denmark as well[62], so the dataset is partly “contaminated”, but even so there is still a remarkable result to be had from this exercise, and the difference would simply be expected to be even clearer had the Danish log been free from English queries.

The L^{NZ} log consisted of 126,798 queries issued during the first six months of 2012. Before the analysis could be made, 36% of the queries had to be removed from consideration because they were non-linguistic; one third of the users simply search directly for document numbers and some use other codes such as project numbers and author initials. Other than that, the dataset is remarkably clean due to its enterprise-internal origin. A query length analysis on all of the non-excluded L^{NZ} queries reveals a result that is once again identical to that from the other studies (fig. 3.17). Thus, even in an intranet environment, this distribution holds.

[62] As much as 41% of the words occurring in the Danish log *may* be interpreted as English words, but as many of them are technical terms, these may just as well be considered Danish loan words.

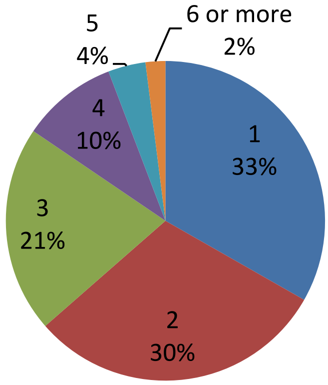


Fig. 3.17. Distribution of query lengths in the entire L^{NZ} as measured by the number of terms in a query.

However, if the log is split into the Danish and the international (non-Danish) sections, an interesting change occurs (fig. 3.18-3.19).

Number of terms	Total	Denmark	Other countries
1	33.2 %	36.7 %	28.8 %
2	30.3 %	30.1 %	30.6 %
3	20.9 %	19.6 %	22.6 %
4	9.7 %	8.8 %	10.9 %
5	3.8 %	3.1 %	4.7 %
6 or more	2.0 %	1.6 %	2.5 %
Total	**100 %**	**100 %**	**100 %**

Fig. 3.18. Query length in L^{NZ} as measured in number of terms. The countries contributing most heavily to the non-Danish log are USA, China, Brazil, Japan, India, Indonesia, and Switzerland.

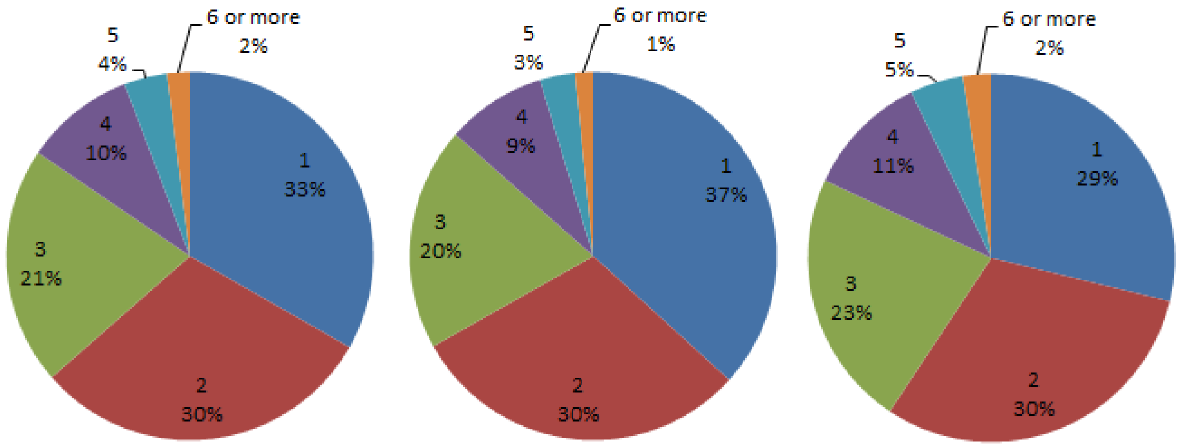


Fig. 3.19. Query length in the entire L^{NZ}, the Danish section and the non-Danish section, respectively.

Apparently, the standard distribution arrived at when the entire mixed log is analysed is composed of sections with varying characteristics. The Danish log has a much greater proportion of 1-word queries, as expected, while the non-Danish section has a correspondingly lower proportion of such queries.

It is possible that this mixed origin of distributions that results in the **standard query length distribution** found with all of these query logs except L^{WID} is also present in the other cases. After all, they are all likely to contain a mixture of languages, including English and Danish or similar languages. Perhaps it is this mixture of language types and uses of the tool that when combined together produce the standard distribution. This might even explain some of the aberrancy of L^{WID}, as most queries in this log are in English, in spite of the relatively large amount of variation, and this would tend to favour a greater amount of 1-word queries.

3.2.2 Query efficiency

As longer queries have been shown to result in the best matches (Buckley, Salton, & Allan, 1994), it is easy to think with Kelly *et al.* (2005, p. 457) that "there is an apparent mismatch between what interfaces encourage users to do, what users are doing and what has been demonstrated to result in good retrieval". Indeed, it has become popular among system designers and IR researchers to employ the method known as **query expansion**, i.e. automatically adding words to a query in order to improve the result, an idea which is based upon the assumption that more terms mean better results. However, this very assumption has been questioned by Ruthven (2003), but confirmed by Kelly, Dollu and Fu (2005), and consequently remains controversial. Although Belkin and his colleagues found that the users who entered the longest queries were more satisfied with the results, they also noted that it did not in fact improve the probability of locating correct information (Belkin N. , et al., 2003)[63]. Azzopardi (2009) investigated the relation between query length and the resulting increase in precision and concluded that the most efficient queries are in fact those of two to five words. Furthermore, what he considered to be the most natural types of queries tested in his study are optimally efficient when they contain no more than three terms. It would seem, therefore, that it is not without reason that users rarely provide longer queries; it simply gives maximum value for their efforts. Furthermore, it is to be noted that one line of research has recently been developing sophisticated methods for reducing long queries to shorter keyword-based ones, thus effectively counteracting earlier attempts at persuading users to be more verbose (Stein & Hagen, 2011). However, web search queries are also

[63] Interestingly, almost 45% of answers were deemed incorrect by the research team, irrespective of query length, and this is a result also arrived at by Wang *et al.* (2000). If this indicates that almost half of the information that people find on the internet leads them to draw the wrong conclusions, that is surely a serious flaw of information transfer.

reported to be in the process of becoming longer. The need to reduce queries may of course not have arisen because long queries are inherently ineffective, but simply because search engine developers have been focusing so intensely on dealing with short queries that current systems are not able to cope effectively with longer ones.

Azzopardi's study concerns web search; in enterprise search, which is our main focus here, it has been suggested that users may be "more willing than, say, average web search engine users, to express their information need in a more elaborate form than by means of a few key words" (Balog, Weerkamp, & de Rijke, 2008). Consequently, the mean query length might be supposed to be slightly greater as compared to web search. However, as enterprise search users are also within a much more clearly defined domain, the need to disambiguate terms may be correspondingly reduced. This seems in fact to be the stronger effect in the examined logs, with their average query lengths of 1.54 (L^{TDC}) and 1.37 (L^{WID}) words, respectively.

It has been found that familiarity with the topic has an effect on query length (Belkin N. , et al., 2003); people with a high degree of uncertainty about what they are looking for generally employ short queries. Encouraging such users to provide more query terms may thus not be advisable. True, longer queries seem to be associated with a reduced need for going through many iterations with modifications of the query, but on the other hand longer sessions may be exactly what such insecure users need. The users of the Workindenmark site seem to be particularly insecure and uncertain about what to search for, which is hardly surprising given that they are looking for jobs in a country they do not know very well and often in a language that they do not master. Perhaps familiarity with the topic is another of those situations where the ideal system would adapt to the characteristics of the user and behave accordingly.

3.2.3 Query length and task type

It was mentioned in passing above that apart from the difference in language types, the typical tasks associated with a certain LAT might also have an effect on query length. It will be useful here to anticipate some of the results from Part III where it is shown that queries differ in what aspect of the user's needs they refer to. Some refer to the initial problematic stage in which users find themselves while others refer to the ideal stage that they want to reach; some refer to topics and others to resources that the user wants to locate, etc. It would be natural for such differences to be reflected in query length variation, and this is indeed borne out by the data.

L^{WID} was manually analysed according to these referential differences – or **targets** as we shall call them – and thus query lengths can be calculated for each of the main types represented in the log (fig. 3.20).

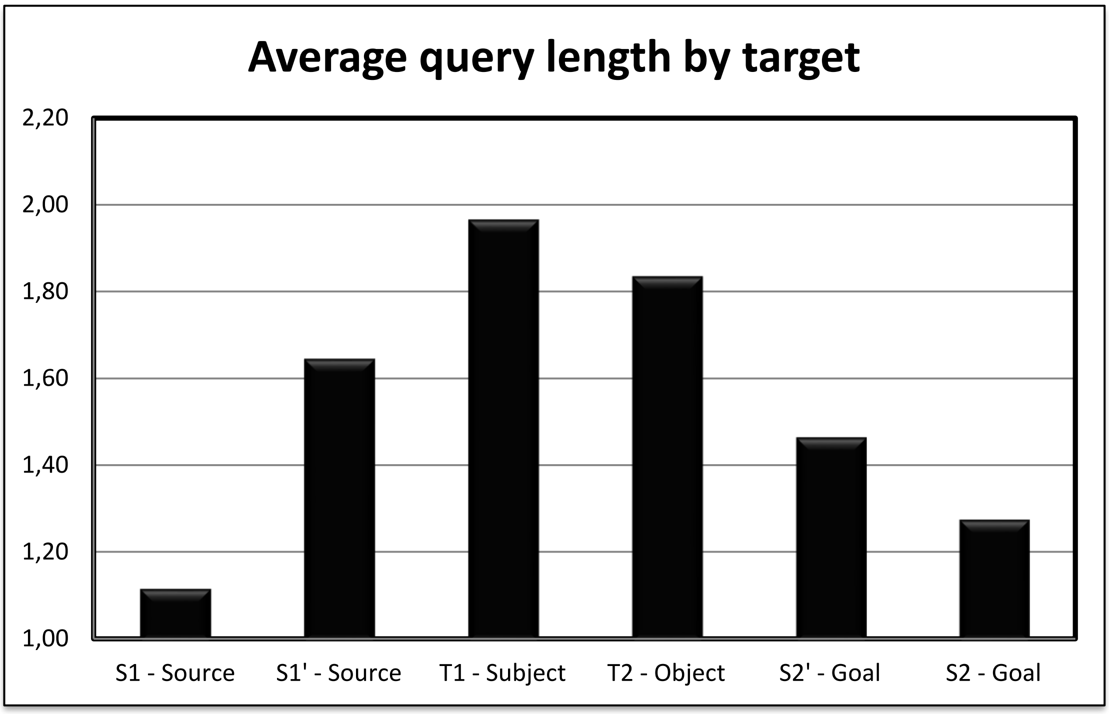


Fig. 3.20. Variations in query length in L^{WID} by query target. Source: The initial troubled situation. Subject: A topic on which the user wants information. Object: References to text to be located. Goal: The desired outcome. See Part III for details. Note that the dataset contains far more queries of the Source and Goal types than of the Subject and Object type. The numbers of queries as extrapolated from the analysed log reduced to 10% are: Source 5,062.3 x10; Goal 12,630.1 x10; Subject 542.5 x10; Object 269.2 x10.

There can be no doubt about the difference in query length among these types; it is clear that whether the initial situation or the desired outcome is described, the query is equally short, while searching for information on certain topics or anticipating stretches of text to be located implies longer queries. One may be surprised to find that topical queries are the longest in this dataset, but only until it is realized that typical topical queries in L^{WID} include the following.

3.5	**salary of nurses in Denmark**	WID (1)
3.6	**ulykkesforsikring i forbindelse med virksomhedspraktik** 'accident insurance in connection with company internship'	WID (1)

Incidentally, it should be noted once again that we do not count phrases or even terms, but words; a query term often consists of more than one word, especially in a language such as English but also in Danish. Furthermore, there is a relatively large amount of complex queries that incorporate references to more than one target, and these naturally consist of much longer queries, having an average of 2.55 words (the extrapolated number of these is 1,143.2 x10 queries). Thus, it may be concluded that the types of task typically performed with a certain tool do indeed have a significant effect on the average query length, and that even when the linguistic makeup is similar, only tools with mixed types of uses should be expected to conform to the Standard Query Length Distribution. The fact that L^{WID} has an unusual distribution of query types with a very small number of topical queries and a large number of Source- and (especially) Goal-targeting queries (see Part III) which tend to be shorter thus suggests an additional explanation for the deviating average query length in this case; the standard query length applies to sites with a mixed set of uses. It is these different uses that we shall be exploring in Part III in order that sites with different sets of uses may be designed in suitable ways.

3.3 User friendliness and unfriendly users – Consequences for tool design

According to Relevance Theory, conveying a message necessarily also implies transmitting the intention to convey the message (Wilson & Sperber, 2004). This is achieved by producing an *ostensive stimulus*, such as a piece of language, which in itself generates an expectation of a relevant input. Wilson and Sperber suggest that an utterance always conveys the assumption of its own *optimal relevance*. Therefore, the hearer is bound to be wondering in which way it might be relevant. In search, the ostensive stimulus consists in the clicking of a search button, whereby the intention to convey the message that he or she wants the LAT to respond in a relevant way to the request is communicated. However, requests come in many shades. Most existing LATs only allow the user to express one function, which consequently must be interpreted in a very broad way, i.e. something akin to *FETCH ME SOME INFORMATION ON X*. Even so, users have adopted the technology for use in many different situations and have evolved ways of exploiting LATs to satisfy highly divergent needs. Accordingly, what correspond to different sentence functions have evolved, even though the LATs used do not supply any way of conveying these more fine-grained intentions. Some common functions in search are those where the query can be paraphrased as *FETCH ME A SITE WITH X IN ITS TITLE OR URL*, *SHOW ME ANYTHING WITH X IN IT*, *HELP ME FIND OUT WHAT X MEANS*, and *FETCH ME SOME INFORMATION ON THE TOPIC OF X*, but there are many others. Already Cooper (1971) was on the trail of the fact that there are such differences in function when he noted that a question such as *Which elements are halogen elements?* might not only be answered with the names of such elements, but also with statements such as *the element whose nucleus contains exactly six protons*. This is a case of the difference between fact-finding and understanding of a term. The problem remains unsolved but is a promising field for research in the diversity approach that comes under the suggested title of *Dynamic IR*.

The request *TELL ME SOMETHING ABOUT THE CAPITAL OF GERMANY* is topically and not factually motivated. If we abstract from the different format of the information required in order to answer such requests, factual and topical motivations are rather similar. The topical request would be interpreted in discourse as *I WANT YOU TO TELL ME SOMETHING ABOUT THE CAPITAL OF GERMANY*. Because the possible interaction with a LAT is so much more restricted than discourse with a human interlocutor, we can reduce expressions even further – the factual one to *the capital of Germany?*, and the informational one to *the capital of Germany*. People do not usually make a distinction between these two communicative intentions or aims; in fact, unless they are specifically designed to understand natural language questions, LATs do not currently incorporate a request type corresponding to the interrogative of discourse[64]. In principle users might perhaps be taught to (sometimes) make a distinction so that the use of a question mark would mean a request for a specific fact, paralleling the use in natural language. This would allow the LAT to react differently, providing a factual answer to the question when a question mark is included and documents on the given topic when no such marking is employed. Another possibility would be to provide separate buttons[65]. In some cases, the LAT might hazard a guess as to which motivation type is the most likely, based on contextual information, but in the long run, the best result might be achieved by allowing the user to convey the function of the request, either before or after the search is performed, either directly or by providing diverse results. As it is, a query such as *the capital of Germany* is probably much more likely to produce responses in the form of documents about the capital rather than actual answers to the question, which are only incidentally present.

3.3.1 Exploration tools

It is well known by now that users do not usually have clearly defined information needs before the initial search, and that they generally experience severe difficulties in stating their own needs (or, in other words, converting a task into information needs).

[64] At least, traditional search engines are not geared to understand such queries; surprisingly many people do enter questions (cf. section 9.3.2), but they are not likely to get many results unless the questions have specifically been included in an FAQ document (Verberne, Boves, & Kraaij, 2011). It is interesting to note that even though by far most queries are of the topical rather than the factual type (i.e. *TELL ME ABOUT* rather than *I ASK YOU* requests), early researchers largely focused on determining relevance in the factual case. Thus, good responses were supposed to either answer exact information needs or help the user reach certain conclusions (Wilson P. , 1973). This emphasis stems from the seemingly harmless – but erroneous – notion that search is a type of question posing (Sparck Jones, 1989), and also the enthusiasm with artificial intelligence that was current for some time until it was realized that the ultimate goals of such work were neither possible nor desirable. In fact, informational search has more in common with imperatives than with interrogatives. This means that users' needs are much more vague than believed by these early authors.

[65] Lexicographers at Aarhus University have experimented with this kind of approach in their dictionaries at www.ordbogen.com, and the results seem promising. Various buttons may be specialized for certain common task types.

> When users approach an information access system they often have only a fuzzy understanding of how they can achieve their goals. Thus the user interface should aid in the understanding and expression of information needs. It should also help users formulate their queries, select among available information sources, understand search results, and keep track of the progress of their search. (Hearst, 1999, p. 257)

It was insights such as these which already made Doyle (1963) suggest that an information retrieval system should provide the user with an efficient means of exploring available information, rather than endeavouring to identify relevant documents. Thus, precisely because search is a process, it would often be better to provide a means of topic exploration, or dialogues that establish user needs, rather than attempt to retrieve relevant documents directly (Swanson, 1977). Swanson mainly envisions a system in which the dialogue allows the user to "ask for and receive progressively increasing amounts of information about a document" (1977, p. 143). Ingwersen (1996, p. 39) suggests that this approach be followed "in the case of an ill-defined or vague intrinsic information need", thus allowing the user to "follow and explore different conceptual paths in the information space". Nowadays, technology has advanced so much further that it is reasonable to propose a system in which the tool takes the initiative of gathering information on the user context.

Something along these lines, though more strictly classifying and less associative, has been proposed in a report by the Delphi Group (2004). They suggest that when designing information systems for finding information among the documents in a company, search should be complemented by a taxonomy that classifies the documents according to topics, the idea being that the very process of browsing a hierarchy would provide the user with some initial awareness and even understanding of the topic through its structure. This would guide the user towards a clearer understanding of his or her real information need. It is perhaps to be feared that a strict taxonomy will be too rigid in many cases, though in some companies it may well prove sufficient and even appropriate for organizing the internal information. In systems intended for external use or covering a more heterogeneous domain than much internal information, less rigid, associative links are probably to be preferred.

The main conclusion to be made from this discussion is that encouraging users to take time developing their thoughts and clarifying their needs should be an integral part of a search system. Information retrieval is not, after all, merely a question of retrieving information, but rather of helping users to reach their goals and satisfy their needs.

> IR techniques will thus be most successful when acting as instruments that support the human retrieval of information. (Ingwersen, 1996, p. 8)

This includes understanding the task properly and going through a process of knowledge acquisition; an IR system could be so much more than just a search algorithm.

3.3.2 Meta-adaptive tools

At the end of the day, the traditional search box still has its merits in that it is very fast and simple to use. Thus, on the basis of the above it would seem that a combination of exploration and traditional search is needed: Exploration in the initial stages and straight-forward search when a focus has been gained.

The empirical study by Navarro-Prieto *et al.* (1999) in particular raises the question of whether IR systems need to work in different ways depending on the user's level of experience. Even when skilled users are able to find what they need, less practiced ones would certainly benefit from improved suggestions and automatic query type recognition. IR researchers themselves presumably tend to be in the former category and may easily come to assume that their way of searching is standard. This is clearly not the case. Search novices typically fail to plan ahead and consequently obtain poor results from systems that would be capable of returning much better responses if properly used. This suggests that interface design should encourage strategic planning and help novices get better results. In other words, the ideal IR system needs to adapt to different users according to their experience when it comes to searching as well as in relation to the task they are facing.

The ability to adapt is a key notion, since no one solution will ever be able to help all users alike. There is so much variability in behaviour and expression that all but the coarsest predictions based on user input must always be futile unless attention is paid to the kind of user that originated the request. In other words, subjectivity is subjective and so adaptation needs to be adaptive.

Expertise in the domain of the requested topic likewise has an important role to play. Users with little knowledge of a topic tend to want rather concise and ordered information; they are looking for answers to relatively specific questions, because such questions are the only entrances they have into the unknown territory of a previously unvisited knowledge domain. On the other hand, people with a greater understanding of the domain in which a term figures tend to prefer longer stretches of text with more background information. Also, they are less afraid of getting lost by following unknown scent paths to informational cul-de-sacs than novices who tend to prefer to be led by the hand. This has been demonstrated with *adaptive hypermedia* in cases such as e-learning; beginners prefer the linear step-by-step introduction to a subject, whereas the more experienced are keen to follow suggested links that may lead them away from the marked-out path (Brusilovsky, 2003). The point to note is that needs are highly variable, almost to the point of making it impossible to construct a tool that will serve all users. A system not only has to be adaptable to the current user, but also to have a repertory of different adaptation strategies for different types of users, resulting in what Brusilovsky has called *meta-adaptation*. Adaptive hypermedia systems build a model of the user, incorporating his or her preferences, goals and knowledge at a specific point in time, and draw on this profile throughout the interaction with the user with the purpose of either providing suggestions as to where to go next or to actively hiding links in order not to confuse the user. This is especially useful in e-learning, where the user's knowledge within the confines of a certain subject can be defined with relative confidence. It is much more difficult in the context of a more general LAT. Even within e-learning, however, Brusilovsky emphasizes that any one adaptive hypermedia technology is never going to fit all users, and that consequently there is an urgent need for what he calls *meta-adaptive hypermedia systems*. Such a system would choose the appropriate adaptation strategy based on its knowledge of the user.

3.3.3 Accessing the user's mind

Adapting an IR system to the user requires access to information on user characteristics and preferences. There are various ways of obtaining such evidence, and even though Cooper (1971, p. 32) felt – quite reasonably – that "no system has direct access to a user's mind", and that consequently no system could know what a user truly needs, it is not the case that no such information can be obtained. There is no reason to give up on trying to adapt to the user's context.

As pointed out by Xiang *et al.* (2010, p. 451), "the absence of context information in document ranking models is probably partially due to the difficulty of obtaining context information". However, there is a range of available methods for gathering contextual information on the user:

- **Dialogue**: Asking the user to provide the information
- **Post-filtering**: Allowing the user to interact with the system after the search, thereby incidentally providing the desired information for storage and later use
- **Inference**: Inferring characteristics and preferences from user behaviour
- **Crowd-sourcing**: Inferring characteristics and preferences from other, similar users

Furthermore, information may be elicited at every search session, or stored in a dynamic user profile, or as part of a login profile filled in by the user.

These methods are all useful, but for different purposes. Clearly, crowd-sourcing is not normally appropriate for deciding whether a document is sufficiently recent, for example, and starting a dialogue about what sources the user would trust before any results have been retrieved would be extremely tedious for the participating user. In such cases, post-filtering is far more suitable. It is helpful in this connection to distinguish between different types of user information:

- Demographically and geographically dependent preferences such as language preferences and age suitability
- Device-dependent preferences such as preferred formats
- Spatiotemporal information necessary for the calculation of recency and geographical suitability
- Personal preferences such as preferred document length, trusted sources and required recency
- Domain and expertise information such as the user's previous knowledge and interests
- Task information such as the overall goal of the search session and the stage reached

Note that these types only refer to user preferences and not to the characteristics of resources that will make a certain resource relevant to a person with those preferences. Such information must be gathered by a separate process analysing documents.

With the finer distinctions of types of preferences suggested here, specific recommendations can be given in relation to each. In order to reduce the necessary filtering left to the user, demographic and device information may be stored in a login profile. Spatiotemporal information should ideally be collected continuously in the background, especially with mobile devices. Personal preferences could be obtained directly from the user, either by elicitation or post-filtering, of which the latter method is probably to be preferred in most cases. Domain and expertise information could be stored in the login profile, supplemented with a dynamic user profile that registers the various changing interests over time. Task information can be accessed by direct elicitation, or possibly by post-filtering in some cases.

An alternative to all of these ways of gathering information is simply to present different sets of results side by side and let the user choose from the collection. This is likely to be a very effective solution. However, only a very few sets can be aligned in this way, so finer distinctions must be made by employing some of the other methods.

⬤ Post-filtering

As noted above, preferences may be dealt with either by asking the user beforehand, preferably as part of a personal profile that is activated through a login, or after the results are shown, in the form of a filtering feature that lets the user remove items according to certain criteria such as language, style, format, length, source, authoring date, and place of origin. As pointed out by Jacsó (2006, p. 73), post-filtering is often the better option, because users "do not want (and can hardly be expected) to make commitments prior to seeing what their topical search without any limitation would bring up". Thus, users are typically both unable and unwilling to state their preferences in advance without knowing whether they will benefit from doing so. Consequently, unless there is some stored profile data about the user, preferences are best dealt with after the presentation of results. For each result presented it might for instance be possible to remove all results from the same source, of the same format and so on. This would enable users to quickly trim down the top results to a manageable set, even if there are sources that they do not trust or cannot easily read and so on.

Some preference parameters may simply be pre-set because users of a certain site are expected to be interested in resources of a certain kind. Alternatively, a site or a search engine may allow users to specify their preferences without a login in the form of a predefined set of standard user types. This method is employed in some web pages, where it is possible to access the information from different points of view depending on user type. It is not normally exploited in search, however. Even in sites that implement user types as part of the site access structure, such as the site of the Danish Physiotherapists, *fysio.dk*, the employed search system bypasses the choice of user type completely and ignores it when retrieving and ranking documents.

⬤ Encouraging diversity

Since we do not know exactly what users want, it might well be safer to include as diverse results as possible among the top-ranked results, instead of just filling the top 10 with documents that all reflect what is believed to be the most likely interpretation of the query. In this way, everyone is likely to find something interesting, even if the amount of noise is also relatively high. The idea of **document diversity** has been taken over from risk theory in economics where putting all of your eggs into one basket is unthinkable[66]. Another notion

[66] Incidentally, it would also counterbalance the effect of the so-called *filter bubble* (Pariser, 2011) – the ethically dangerous consequence of giving users too much of what they want and too little of what they prefer not to see.

taken from this field is **portfolio theory**, the idea being that ranking is not just a question of finding the right documents, but the right *combination* of documents (Wang & Zhu, 2009). The results should together provide the user with the necessary information, rather than all of them saying the same thing except number 400 which supplies the last, essential piece of information.

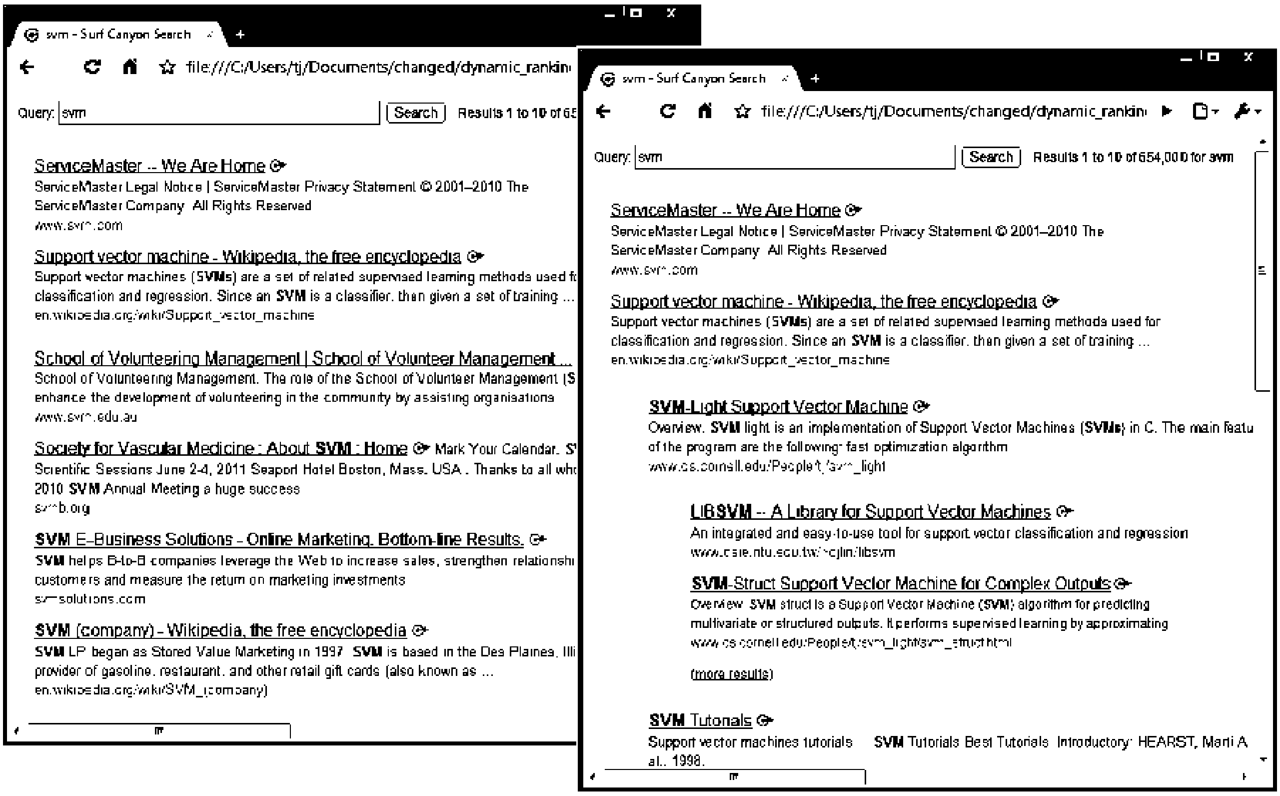


Fig. 3.21. An interactive interface with dynamic ranking as proposed by Brandt et al. (2011). The initial ranked list is maximally diversified so that each result concerns a different sense of the highly ambiguous query term *svm*. The subsequent levels provide more distinctions within the chosen category.

Document diversity retrieval has become the new trend, and has a great potential for optimizing the likelihood of finding something relevant. Even if none of the particular documents shown on the first result page is a perfect match, this kind of strategy allows for more directed interactive features. Thus, it has recently been suggested that dynamic interfaces be made in which the user is allowed to click on the result that seems to match best, which would then cause that node to expand into a tree of new "subtopics" that the user could choose from, and so on. This would be especially useful in the case of semantic ambiguity, where the first step could be used to disambiguate the sense, whereas the next step would be more about the aspect of that concept that was sought, or the particular task the user is facing, etc. (fig. 3.21). The disadvantage of such a system is of course that it requires the construction and maintenance of a huge ontology in order that the system may know what documents belong in which subcategory. Also, it relies on the use of a semantic web where the exact sense of a word is known by the system (or an alternative method of determining word semantics from context).

In a way, document diversity means giving up on the attempt to reconstruct user intentions. Though this may be hard to accept, there is a sound logic behind the idea – especially in multi-domain tools such as web search engines. In smaller domains, diversity maximization is likely to be less effective, though perhaps not to the point of being useless. Of course, being able to present diversified results in itself hinges on the ability to recognize differences, and so an improved understanding of possible intents related to various types of document is requisite even for this kind of system.

Much work has been invested in designing user-friendly search systems, but the main problem is that users themselves typically assume that search is indeed a mere question of collecting information and hence attempt to circumvent the process leading up to that stage. As people have become accustomed to traditional search boxes and expect to receive a list of results directly, they frequently bypass the initial phase of exploration, which in itself reduces the chances of success severely. Therefore, it may be worth trying to improve search engines' ability to interpret queries and identify relevant documents without recourse to exploratory methods. However, the best solution would probably be a kind of combined interface in the spirit of Dynamic IR in which the user is never forced to use one or the other method but always has a choice, yet one which encourages the user to explore the topic in the initial stages and to focus increasingly on a specific kind of information. In this way, users could choose whether they want to lead a dialogue with the tool, which might help them find something truly helpful, or whether they just want to do it the fast way and risk receiving low-quality responses of informational fast-food.

Unfortunately, as pointed out by Azzopardi (2009), empirical studies have not only found that users profess to desire explicit relevance feedback features that let them choose the exact ways in which the query should be expanded instead of automatic mechanisms hidden from view, but also that when given such features they do not in fact take advantage of them. However, it has also been indicated that user behaviour may be changing, and that users may be becoming more comfortable with providing feedback (Anick, 2003).

The studies by Azzopardi and others show that most people generally do not want to apply filters or answer questions about the nature of their needs before the search. What has not to my knowledge been tested is what the response would be to applying such filters *after* the results have been shown in order to reduce the amount of results shown by removing irrelevant ones. This is likely to be a much more fruitful approach, which is completely in line with the recent developments in Dynamic IR. Once the results have been returned, the user may be given the choice to remove some sets of results by disambiguating the sense of terms, by stating that a suggested resource is too old and hence only newer results should be shown, by removing resources from a distrusted source, texts written in certain languages and resources of certain formats and types. These criteria all contribute to the reduction of irrelevance, and we shall see in Part II that irrelevance is often an even more important concept than relevance itself. However, some functionalities hinging on *relevance* might also be added, such as the ability to get more results that are similar in some way to the one that has just been seen – either topically or with regard to its type and origin, for instance.

3.3.4 Moving on

A LAT must on the one hand compare a query to the documents in the collection and on the other match the document topics or semantic contents to the user's interests and preferences. Traditionally, only the first part is automated, while the second is left to user perusal of results. In principle, this is an ingenious solution, as no system is better or more efficient at solving the task than a human observer. The reason that search tools should be built on an awareness of cognitive issues is thus not that they should replace human cognition but that they should enhance it and certainly must involve it.

> User Model Building is consequently mainly applied (or minimized) to *infer adequate backing* of the user, in order to make maximum use of *human intelligence*. (Ingwersen, 1992, p. 205; emphasis original)

However, there are limits to the capacity of the human mind, and certainly to the time users will spend checking results. Consequently, much effort is put into trying to refine the results as much as possible before handing them over to the critical eyes of the user. Few users have the patience to check many pages of results for useful material (see section 10.1.4 for empirical results).

Limiting the amount of results returned to something that the user can handle can be done in several ways. One is to improve the system's "understanding" of queries in the first place[67]; this is the solution

[67] It is often claimed that "IR systems and their interfaces cannot 'understand', and will never come fully to understand, user requests and document texts" (Ingwersen & Järvelin, 2005a, p. 161). This of course depends on what one means by *understanding*; the system collects information and, based on this information, is able to decide between a set of possible predefined actions, and this, the present author thinks, is a kind of understanding. The fact that such understanding is much less complex than human understanding derives from our much greater capacity to associate ideas and is related to the complexity of our coding capabilities as discussed by Maturana (1980), cf. section 4.6.3.

that we shall attempt to work on here. Another is to collect more information on the background and context. This is advantageous in some instances, such as mobile applications, but in most cases the amount of information that it is possible to gather in this way without being intrusive to the search process is highly limited. A third way currently explored in what I have called *dynamic IR* is to improve the user interface in ways that let the user interact dynamically with the tool, thus combining the cognitive powers of both in the quest of identifying the exact information that is useful in the task at hand (Albakour, et al., 2011). It is probable that a combination of all three methods will yield the best results.

It is immediately apparent from the preceding chapters that there is much more to searching than simply matching a query to a document collection. Rather, it is likely that a major part of the problem of getting useful results actually resides not in the system's performance as such but in the preliminary stages of coming up with the right kind of query, and even grasping the structure of the topic involved. Thus, search engines would certainly do well to attempt to help the user through this demanding process of knowledge acquisition instead of focusing solely on providing a correct match (Ingwersen & Järvelin, 2005a). This was in fact one of the most important insights of cognitive IR in the early 1990s, but so far it has not caught on among the web-based search systems. In the remaining two parts of this thesis, we shall be looking closely at the choices made by users during the search process. Hopefully, this will contribute to furthering progress in that direction.

Part II

The need for relevance

Chapters 4-6

4 Relevant information

> *What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.* *(Simon H. , 1971)*

> *We are still in search of a theory of relevance applicable to the context of information science and particularly IR. In other words, we are still in search of a conceptual basis, a set of testable principles and propositions, to explain the notion of relevance applicable to information science practice, to explain its manifestation, and to predict its behavior and effects.* *(Saracevic, 2007)*

Ever since the formation of information science as a distinct discipline in the 1940s, **relevance** has consistently been considered its most fundamental and central notion, or as Tefko Saracevic put it:

> [I]nformation characterized by its relevance became the key notion in information science. And the key headache. (Saracevic, 1996)

Information retrieval is fundamentally dependent on the concept, and the literature abounds with descriptions of its effects, but in spite of being a seemingly straightforward principle effortlessly adhered to by everyone, this remarkably slippery phenomenon is also one that it is exceedingly hard to pin down and define (Saracevic, 1975). In fact, it has been described in so many seemingly incompatible ways that it has been very difficult to come to any kind of agreement on an understanding of the term, which has during the last three quarters of a century been studied by philosophers, logicians, discourse analysts, computational linguists, librarians, psychologists, neurobiologists and psychophysicists among others, with varying goals and consequently with divergent results.

For a long time, search engines were held to be merely tools for retrieving information on a given topic. This was a notion that had grown out of library system research along with the actual tools and had never been seriously challenged. Even when a category of *usefulness* was added as a necessary ingredient of the concept of relevance, it was still common to consider topicality to be required for a resource to be useful though the opposite was acknowledged not to hold.

> Usefulness is strongly dependent on topical relevance, but usefulness is also dependent on other factors. Topical relevance is necessary but not sufficient for usefulness. (Glover, 2001, p. 78)

In later years, however, information retrieval researchers have increasingly come to recognize search engines for what they are – or rather, what they ought to be: A tool for helping users acquire the information they need in order to fulfil a task.

> The ultimate goal of information retrieval (IR) research is to create ways to support humans to better access information in order to better carry out their (work) tasks. (Järvelin, 2011)

This new goal has resulted in a focus on how to identify truly useful information as opposed to information which is merely on the requested topic. It is by now a well-documented fact that exhaustive information on a topic is neither necessary nor sufficient for the satisfaction of a user's needs (Harter, 1992, p. 603). However, exactly what is required is a rather more vexed question. Users and their needs differ greatly across searches, and determining the intentions of a user, or even assigning users to user groups with similar general needs, can be exceedingly difficult. One solution has been to study the contexts in which needs develop and attempt to capture some of the contextual data in order to be able to provide results which are more specifically adapted to the user's needs. Another has been to classify queries into a range of types according to the kind of intention that they are likely to represent. Whatever solution is chosen, it is imperative to have a powerful theory of relevance that explains why some documents are more relevant than others to a particular user in a given situation.

Originally, relevance was thought to be a simple relationship between a document and a request. Hence, in the now classic experiments performed at Cranfield University in the 1960s, sample collections of documents were selected for tests, along with sets of test queries (Cleverdon, 1960). Experts in the topics treated in the documents were asked to determine which documents were relevant to which queries, the idea being to isolate pure, simple relevance in order to evaluate the effectiveness of different search algorithms. The measures of *precision* and *recall* were introduced in order to quantify the degree to which a given system was able to identify the relevant documents and ignore the rest. However, it soon became apparent that relevance is not an objective matter which can be treated in isolation from users. Rather, users interpret documents differently, and queries may represent a multitude of intentions. Consequently, many researchers felt that a new notion was called for to take the place of relevance as it was then conceived. Suggestions have not been lacking. In fact, there are just about as many interpretations of relevance as there are IR researchers. Furthermore, as noted by Schamber, Eisenberg and Nilan (1990, p. 760), "the many definitions of kinds of relevance often overlap and contradict one another". This has resulted in what Mizzaro (1998) dubbed the *relevance pool* – an unorganized melting pot of terms for different isolated phenomena which do not add up to an understanding of the phenomenon as a whole (fig. 4.1). As a result of this, it is as true today as it was in 1990 that "an enormous body of information science literature is based on work that uses relevance without thoroughly understanding what it means" (Schamber, Eisenberg, & Nilan, 1990, p. 756). As late as 1998, it was still true that "although relevance has been debated for more than three decades, a clear definition or viable operationalization within the context of IR system evaluation has not emerged" (Spink, Greisdorf, & Bateman, 1998). However, intense research during the 1990s paved the way to a better understanding of the notion, and with Borlund (2003) it looked as if a synthesis of the many proposals of the previous decades was finally going to be possible. Today we are able to see relevance as the complex and multifaceted notion it is (Saracevic, 2007). Even so, the phenomenon can hardly be said to be well understood, and a coherent model is still conspicuously absent. In this thesis, we shall attempt to model relevance and for that purpose it shall be defined here as the quality that contributes to a subject's need or desire to attend to a stimulus.

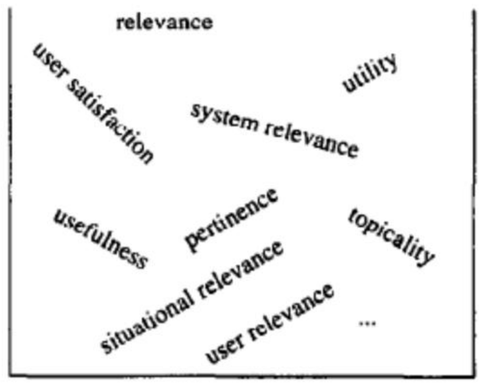


Fig. 4.1. The *relevance pool* (Mizzaro, 1998). The illustration only includes a tiny selection of the multitude of terms for relevance phenomena used in the literature.

Relevance is the quality that contributes to a subject's need or desire to attend to a stimulus.

For reasons given later, this is in fact not to be considered a definition of the phenomenon as such, but only as a delimitation of what it is that we shall try to model.

For a long time, relevance judgments were simply treated "as though they emanated from a 'black box,' whose inner workings were of no particular interest as long as the black box did its job, i.e., providing a score that could be used as a criterion measure" (Cuadra & Katter, 1967). The very fact that people tend to put their faith in poetical metaphors and paraphrases like this shows how enigmatic the phenomenon has been considered through the years. In his book on information retrieval interaction, Ingwersen (1992, p. 54) still simply refers to the phenomenon as "the *Dark Matter problem* in IR" and devotes little more than a single page to it, presumably because of not having any idea as to how to deal with it properly. This treatment was repeated by Janes (1994) in the guise of the "Big Black Question Mark" of relevance, and is true of many works even today. Despite this designation, Janes did think that it would eventually be possible to get at the essence of relevance.

> [I]f we have a satisfactory theoretical framework (…) and some reasonable methodologies to study these differences in users as well as in experimental subjects, then it may be possible to begin to pick apart the black box of 'relevance' and see what wonders reside within. (Janes, 1994, p. 168)

Saracevic later made a stab at providing such a framework for understanding relevance that would be compatible with Ingwersen's excellent work on the cognitive aspects of search behaviour. However, in spite of significant advances through the years, a truly satisfactory framework and a corresponding theory of relevance is still very much a desideratum, as even Saracevic himself admits in his later work (witness the quote heading this chapter). It is therefore high time that the voluminous literature on relevance was unified into an integral picture embracing all of the different kinds of relevance and a model built based on the findings. However, gathering all the descriptions of relevance in one place will not satisfy that need; an explanation for the variation must be given, and the model must be able to reflect the diversity while at the same time showing how the various phenomena involved relate to each other and contribute to the final relevance judgment of a resource. So far, nobody has succeeded in identifying what binds the different manifestations of relevance together as a unified whole, in spite of some valiant attempts (Mizzaro, 1998; Saracevic, 2007). Consequently, relevance has remained a poorly understood jumble of different types, each one proposed in order to account for yet another phenomenon which one would also like to count as an element of relevance assessment. Authors have generally focused on introducing new types of relevance which were thought to better explain the particular kind of situations that they were especially interested in, and this persistent contrastive focus on comparing innovative models to the traditional or previous understanding has the unfortunate consequence that new ideas can usually only cope with almost as narrow a sphere of relevance phenomena as the preceding models. Nobody seems to have been able to step back and consider relevance in its entirety so that an understanding can be reached that covers the entire phenomenon rather than the one particular aspect of it which is currently under study. It is this tremendous chasm that we shall attempt to fill below, at a time when countless theories have already been built precariously on top of it. Reference will have to be made to philosophy, linguistics and cognitive science as well as to empirical studies of actual search behaviour.

4.1 Aims and methodology

The main purpose of this second part is to investigate whether relevance in its most general form can be split into its component parts, different combinations of which will explain the various phenomena that the existing theories were conceived in order to cover. We shall find that this is indeed possible. Relevance cannot be turned into a simple concept; it must be considered a highly complex phenomenon, yet exactly for this reason it can be decomposed into a range of factors and relations. The complexity of relevance is explained here as the combined effect of six basic **dimensions of relevance** and an in cognitive terms foundational hierarchy of three main types of **relevant contexts**. The hierarchy is based on the findings of the philosopher Alfred Schütz – whose work was originally introduced to the IR community by Saracevic (1996) – and consists of 1) that which is of current interest or concern, 2) that which is currently considered problematic and in need of a solution, and 3) that which is of general interest given the person’s background. Furthermore, the degree of relevance along the six dimensions is judged **bisubjectively** at two points, i.e. users judge the extent to which resources satisfy a criterion, and also the extent to which the criterion must be satisfied for a resource to be considered relevant to them. At each point, three different **scales** are overlaid on top of each other so that relevance can effectively vary along three sub-dimensions in each main dimension: 1) how much of the resource qualifies, 2) how strongly so, and 3) what are the costs of attending to this relevance dimension. Finally, the resource itself is complex, and it is not given which of the six **facets**[68] of a resource a user is assessing. This adds up to hundreds of different “kinds” of relevance, which are all interconnected and necessary features of the phenomenon, thus explaining the wide variety of manifestations of relevance described in the literature. It is no wonder that mere reliance on empirical experiments of search behaviour has not immediately led to the uncovering of such intricate relationships. All of the involved phenomena are discussed at length below.

The notion of relevance proposed here is at once manifold and unifying; it rests on a complex model that reflects its diverse character, yet at the same time it binds the different aspects together as a whole which serves a clear and important function in human life in general. Unification can only be achieved if all the many aspects are considered and given due attention.

[68] In the sense introduced in chapter 2.

4.1.1 Methodological considerations

The search process is basically an interaction between two participants listening to each other's responses and suggestions: The user and the system. The user-oriented approach to IR that superseded the traditional system-driven approach during the 1980s was the first branch of IR research to take this fact seriously by focusing on "the psychological and behavioural aspects of the communication of desired information between human generator and human user" (Ingwersen, 1992, p. 83). The idea is that studying actual user behaviour will enhance our understanding of general patterns which may help us improve tools so as to fit into such patterns of use and benefit the users. This general approach is one that we shall be adhering to as well. However, we shall take the step to the next level of understanding of human behaviour by adopting the cognitive approach to IR advocated by Ingwersen as a natural extension of the general user-oriented one. This implies that we shall be drawing on the findings of general cognitive research when attempting to understand the mechanisms behind the observed behaviour. Ideally, such behaviour should be recorded on the fly by observing and interviewing searching users. The risk of such obtrusive investigations is that the situation becomes somewhat artificial in that the user is so clearly aware of being observed. The non-intrusive alternative that we shall be using in this thesis is to analyse search logs. The material used for our analyses is drawn from natural non-laboratory search conducted in two different sites powered by Ankiro Enterprise Search.

The last couple of decades have seen the publication of a series of brilliant empirical studies aiming to uncover the principles behind relevance by asking test subjects what criteria they employ in relevance assessment of resources[69]. This has shed some much-needed light on the variety of criteria, but as rightly pointed out by Xu and Chen (2006) such exploratory studies do not really tell us much about the deeper workings of relevance; the number of observed criteria is simply much too large to be truly helpful. The abundance of criteria hinders a basic understanding of the phenomenon behind the assessments because they represent surface variation of little consequence for the basic phenomenon. Instead, an independent, coherent theory is needed; one that has been developed by logical reasoning and can subsequently be tested against this material. Such a theory is suggested below, and this in turn introduces order on the data sets and explains the observations that until then were merely observations. What does it tell us, for instance, that users prefer documents written in a clear style? Being well written can hardly be a direct criterion of relevance. It only makes sense if a general theory of relevance is imposed which sees the criterion as a mixture of several observations, each of which is

[69] Excellent examples are Barry (1994), Barry and Schamber (1998), Choi and Rasmussen (2002), and Lopatovska and Mokros (2008).

related to the relevance of the document. The explanation would be that a clear style will help the user understand the contents of the text, and also will make it more desirable to read, and both of these conditions do increase the likelihood of the user finding the document relevant. For this reason, and not the one reported in the study, this criterion contributes to the relevance of such resources. Of course, simply noting down the surface realizations is an important first step in the process of reaching a point at which such a theory can be put together, but it is important to take that step now that we have the necessary data.

The methodology applied here consists in the development of a theoretical framework that explains the effects of relevance judgments reported in the extensive literature and makes sense in a cognitive environment as part of human reasoning and instinct; one which may be tested against the empirical data for fine-tuning as necessary. This is a method that is very common in IR studies (Ingwersen & Järvelin, 2005a, p. 101). Testing can never prove a theory right, but it can falsify untenable claims. If we were to apply a theory claiming that less clarity forced the reader to concentrate harder on the reading task and that this would increase the likelihood of the contents being understood and hence considered relevant in the long run, for instance, that hypothesis would be contradicted by the data since it would predict a negative effect of clarity rather than a positive one as reported. We shall find that the empirical data support the hypotheses underlying the theory developed here rather than contradict it.

4.1.2 Structure of Part II

Being a many-faceted and non-static phenomenon, researchers have had severe difficulties agreeing on a suitable definition of relevance. As a result, many different types or understandings of relevance have been proposed through the years. Section 4.2 summarizes the development of the understanding of relevance and presents the current state of the art. The traditional treatment of the notion and its historical development from librarians' concept of *aboutness* and *topicality* to the fully-fledged subjective and dynamic notion adopted in modern information retrieval is reviewed, and the state-of-the-art *Stratified Model* of relevance manifestations proposed by Saracevic is presented. The *Regulated Flux* model proposed here is based on some of the principles advanced by him, but some misconceptions have been carried over into his and similar models from the long tradition of relevance research and these need to be avoided in order to make real progress.

The phenomenon of relevance is difficult to grasp for several reasons. Section 4.3 presents a selection of them. The section begins with a discussion of whether there is such a thing as *objective relevance*, and if so, what kind of relations such a notion would involve. In spite of the intensive efforts invested through the years in demonstrating how fundamentally subjective relevance is, the objective types tend to persist in relevance typologies as foundational forms of which the others are considered extensions. This is unfortunate, since these supposedly basic types of relevance are in the present author's opinion not really examples of relevance at all, but represent mere proxies.

More than anything, relevance is about the exchange of *information*. Exactly what information is and what makes something informative is a matter of debate, and we shall define our standpoint in this regard in section 4.4 before embarking on a discussion of cognitive issues which hinge on such a definition.

Section 4.5 looks into the cognitive processing and avoidance of information and the *topics* that organize it. Topics enjoy pride of place in the traditional understanding of relevance, yet like that concept the notion of *topic* is in fact highly abstract and not at all well understood. Even defining what topics are is not straightforward. They are generally thought to summarize the subject of a stretch of text, but do they only cover the concepts that name them, or is the entire domain of related issues included in a topic? We shall propose a subjective understanding of the notion based on relevance itself, thus turning the cause-and-effect chain around.

The fundamental human cognitive ability to filter out irrelevant stimuli is likely to be at the heart of relevance and consequently no understanding of the notion can be achieved without a study of *irrelevance* (section 4.6). As human beings, we go through the world with varying immediate interests depending on our short-term goals, but we also have long-term interests that stem from our backgrounds and accumulated experiences. These basic differences reach into search behaviour, and an understanding of them is thus crucial. We shall advocate an understanding of relevance based on the fundamental need to avoid paying attention to the irrelevant.

Chapter 5 deals with some theoretical linguistic issues relating to the production of queries. All language use involves choosing a perspective and a scale on which to portray what the speaker has in mind, and this may have profound consequences for the interpretation, even of queries with their limited contextual clues. The chapter discusses the cybernetics of search and introduces the notion of bisubjectivity as a general principle involved in relevance assessment as well as language use.

Chapter 6 introduces a new understanding of relevance which combines insights from many sources into an integral whole. The new *Regulated Flux* model is presented and a definition of relevance based on the findings is attempted. Furthermore, relevance grading is discussed in the light of the notion of *bisubjectivity* introduced in the previous chapter. Finally, the Faceted Stimulus model developed in Part I and the Regulated Flux model are combined and the collection of contextual information is reconsidered on this background.

4.2 An elusive concept

The notion of relevance is inherent in all scientific work and was hovering just out of reach throughout the development of early science, always there but never actually mentioned, for the simple reason that there was no word for it. In the legal studies, the notion (mostly under the name of *relevancy*) seems to go back to James Thayer (1898), while in the information sciences Samuel C. Bradford is thought to have been the first to use the term *relevant* in an information science context when he wrote of articles as being relevant to a subject in the 1930s and 1940s (Saracevic, 1975). This was followed by the pioneering development of IR systems in the 40s and 50s, mostly aimed at expert librarians rather than being tailored to suit the users. In the 50s and 60s there was a growing feeling that such systems needed to be tested and evaluated as to their ultimate usefulness to users, leading to the famous experiments at Cranfield University in which sample sets of documents were selected along with pre-defined queries and experts in the topics treated were asked to determine which documents were relevant to which queries (Cleverdon, 1960). This method is problematic in many ways and has been duly criticized through the years, but nevertheless is still being used, most notably at the annual TREC (Text Retrieval Conference) workshops in USA, with counterparts in Europe and Japan. It has repeatedly been pointed out that expert evaluations of relevance are not likely to be representative of user judgments, and that the dynamic quality of search is bypassed by such experiments, among other drawbacks. However, in spite of its serious weaknesses, the method has advanced the field in significant respects and is especially valuable in light of the lack of feasible alternatives.

A large amount of studies in relevance and its potential applications to information systems appeared between 1959 and 1976. People were eager to experiment with new ideas and highly optimistic about the potential of computers. Among the more influential papers one ought to mention Cooper (1971) who envisaged an inferential system able to answer questions directly. The period was rounded off by Tefko Saracevic who in a series of papers in the years 1970-1976 reviewed, summarized and evaluated the past research. His 1975 paper was particularly influential and became a basis for much of the research in the following period (Saracevic, 1975).

Maybe this feeling of a completed collective research project was responsible for a somewhat lowered interest in the subject in the 1980s, or perhaps the optimism received a blow when it turned out that the ambitious plans for new information systems were not easily realizable. Whatever the reasons, there seem to be fewer influential high-quality contributions in this period. What did get into print was a mixture of claims that relevance is too subjective to be modelled and backward-looking attempts to hold onto objective knowledge and adjust the expectations instead. Swanson (1986) is a typical ambiguous contribution in this regard; he recognizes the subjectivity of relevance but insists that there is an objective aspect to it that is useful.

In the 1990s, optimism was renewed as information systems underwent rapid evolution and what had been impossible only a few years before suddenly seemed to come within reach. The idea of objective relevance was now more or less abandoned, and a rich literature on the many ways in which relevance is subjective sprang up instead. This period began as the 80s were coming to an end, and really speeded up with the influential paper by Schamber, Eisenberg and Nilan (1990). Also stimulating was Stephen Harter's (1992) paper on psychological relevance with lengthy examples of its application in IR.

A new focus on psychological issues in information studies had begun in the 1980s with papers such as Dervin and Dewdney (1986) who founded the *sensemaking* paradigm, and continued up through the 90s where *cognitive studies* became fashionable. Important exponents are Peter Ingwersen (1992), Carol Kuhlthau (1993) and Ruth Morris (1994). Within relevance research, this period was characterized by the search for the various reasons for subjectivity in relevance and a contextualization of the search through application of notions such as *utility*, *situational relevance* and *pertinence*. This was complemented by a series of relevance assessment studies where users were asked to specify their reasons for deeming a given document relevant. This had been done to some extent earlier, but important contributions such as Barry (1994), Bruce (1994) and Barry and Schamber (1998) established a foundation on which to build. The 1990s also brought reviews of previous research such as Schamber, Eisenberg and Nilan (1990) and Stefano Mizzaro's (1997) paper. However, their conclusions amount to the fact that the notion of *relevance* was still not properly understood and had to be researched further. Consequently, they did not provide milestones in the way that Saracevic's (1975) review had done.

Saracevic (1996; 1997a) attempted to establish a new framework based on previous research and his own insights; he considered the many views on relevance to represent different aspects of the phenomenon and grouped them as different *manifestations* of relevance. This was an important step forward and meant that a new milestone had been reached – a new foundation on which researchers of the following decades could base their work. However, unlike many of his readers, Saracevic fully realized the shortcomings and incomplete nature of his own work.

> [T]he stratified model (…) has not yet enough details for experimentation and verification. It has yet to be tested in a larger interaction study. Clearly, much more has to be done to bring the model to practical applications. A further general weakness of the stratified model is the same as found in the stratificational models in linguistics

> and communication. Decomposition is not that easy, and depiction of interplays between levels, a critical aspect, is hard to specify. (Saracevic, 1997a, p. 317)

As his papers from this period did not represent complete and applicable research but mostly suggestions and ideas, they did not really instigate a new paradigm, although they did inspire many. The same is true of other contributions from this period such as Mizzaro's (1998) framework which did not catch on.

In the present author's opinion, the period in question was not rounded off until Pia Borlund's (2003) paper which reviewed and summarized previous research and also made original contributions. She clarified the feeling of subjectivity in relevance that had been so obviously present but hard to pinpoint during the previous couple of decades. Much more influential was Saracevic's (2007) paper which likewise reviewed relevance research since 1975 and summarized the evolving understanding of relevance. The very fact that this was now possible is notable; earlier on, there had been no real understanding to summarize. Even so, it was clear to Saracevic that relevance was still far from being fully understood and that much more research needed to be done. Even his own ideas from a decade earlier were not so boldly presented anymore, but were recognized to be simply one piece in an immense puzzle.

Several excellent reviews of the literature on relevance have been published at regular intervals throughout the history of the concept[70], but it is still not clear how the many concepts of relevance relate to each other and whether a general principle can be shown to lie behind them all so that an overall model can be brought to include and explain them all. The present contribution builds on the results from the empirical experiments performed since the 90s and on that background extends Saracevic's ideas into a more fully developed and operational model which can be applied in real situations and which explains the phenomenon in a more satisfactory manner.

In parallel with the growing understanding of relevance, another field of research had sprung up; one that investigates how queries can be ascribed to certain types that reflect basic user intentions. This study was more or less founded by Andrei Broder's (2002) very short and provisional paper in which he introduced a distinction between *navigational*, *transactional* and *informational* queries. This trichotomy was to be explored and elaborated on by many subsequent researchers during the following decade and is also the starting point for Part III of this thesis. In this period, research in relevance has been increasingly relegated to conference papers on technical search engine issues as search has

[70] Notably by Saracevic (1975), Schamber, Eisenberg and Nilan (1990), Saracevic (1996), Mizzaro (1997), Borlund (2003), and once again Saracevic (2007).

developed into a billion-dollar market and has been taken up enthusiastically by rapidly growing corporations. This has meant a gradual estrangement from the academic research field, and thus also a reduced interest in a deeper understanding of the fundamental characteristics of relevance. This is potentially problematic, since these researchers often know very little about the cognitive, social and philosophical aspects of relevance which are so important for it all to make sense and be truly useful. This thesis attempts to reconnect contemporary application studies with the academic tradition with a view to allowing future development of useful information systems that will increasingly be able to understand the user's needs and intentions.

4.2.1 From aboutness to situational relevance

The traditional understanding of relevance as a notion in information science evolved in the context of work on improving bibliographical tools in a library environment. In such systems – both before and after the advent of computers – documents were represented by topic labels describing their most important topics, and these subject terms were then compared to the user's query in order to find the best matches. A citation was relevant if and only if it was on the topic of the request. This developed into the so-called *systems view* of information retrieval in which documents whose subject terms match the query are relevant by definition, making relevance an objective category with an indisputable value[71]. Such an interpretation has certainly enabled system designers to test the efficiency of their system in an easy and unequivocal way (Swanson, 1986). However, as pointed out by Barry (1994), this type of system is based on a number of highly contentious assumptions:

1. The assumption that the subject terms assigned to a document adequately describe the contents of that document.

   Assigning topics to documents is very difficult, and there is always the possibility that a given document might also be relevant to some other topic that the classifying librarian did not think of at the moment. When dealing with web documents, the task becomes unviable; all documents cannot be tagged for topics by experts, and often the subjects are so varied and diffuse that even experts would not be able to do it.

[71] This is the conventional way of portraying the development. Hjørland (2010) questions whether such a pure-bred systems view ever existed. He claims that the systems view as it is generally conceived is one conjured up by its very contestants in order to emphasize the greater attendance to subjectivity afforded by competing models. Instead, he suggests that the real difference between the so-called systems view and later models is one of expert judgment vs. actual user assessment in evaluation studies.

2. The assumption that the query is phrased in such a manner as to be representative of the user's information need situation.

This is certainly not necessarily the case; users experience great difficulties in stating what they need, and even if they succeed, it may well turn out that what they actually need is something entirely different from what they thought.

3. The assumption that subject matching of this kind results in the retrieval of documents that the user will find relevant in a specific information needs situation.

It is now known that users are frequently not interested in topical information at all, but rather may have completely different agendas such as finding a certain homepage or performing a transaction. Hence, queries do not necessarily represent topics.

By the 1950s, it was widely recognized that relevance was not a simple question of matching within the system after all. Rather, the user has the last word in the matter. Thus, the notion of **subjective relevance** was born.

Having established that relevance is subjective, the question naturally arose in what ways this might be the case. Determining what external factors influence the relevance of a resource was now considered vital.

> One possible solution to the evaluation problem is to accept that information retrieval systems can achieve, at best, high levels of topicality and should be evaluated on that basis alone. Another alternative, one which would advance the field of information science rather than simply admitting present limitations, would be to explore the possibility of incorporating users' relevance criteria into the retrieval mechanism itself. (Barry, 1994, p. 152)

During the late 60s, possible criteria for subjective relevance were examined, and non-exhaustive lists of as many as 80 different criteria were assembled, based on criteria identified by the early contributors Cuadra and Katter (1967), Rees and Schultz (1967), Cooper (1973; 1971) and Taylor (1968), as summarized by Hjørland (2010)[72]. Another important finding from this period is the fact that there is a crucial difference between relevance to a query and relevance to the information need that prompted

[72] Simply listing the criteria is not very helpful, so we shall resist the temptation of providing the list here. See Hjørland (2010) for details. Later studies have identified many additional criteria.

it; if a document is relevant to the query this does not automatically mean that it is also relevant to the person, and *vice versa* (Harter, 1992). The manifold purposes for which information is required may thus lead to quite different relevance judgments. Furthermore, a document is not simply either relevant or irrelevant in a binary way. Rather, one document may be said to be *more or less* relevant than another (Spink, Greisdorf, & Bateman, 1998; Maglaughlin & Sonnenwald, 2002). Additionally, the availability of highly relevant documents may in itself reduce the relevance of slightly less useful ones, just as the lack of truly useful resources may render a slightly useful document important (Choi & Rasmussen, 2002). Finally, relevance is both subjective and dynamic; since documents inform people in different ways and to different extents, only the user can judge the relevance of a document to his or her need, and the relevance judgment may not even be constant for the same person across the various stages of the search process (Vakkari & Hakala, 2000).

Objective relevance as it was considered under the *systems view* was clearly a fiction, at least when considered from the point of view of searching[73]. However, as can be seen from the date of the quote above (from Barry 1994), the question of how to construct a subjective model of relevance did not find a quick solution. In fact, no single solution has presented itself as a self-evident and convincing model.

The 1960s were followed by two decades of trying to identify the exact differences between relevance as calculated by a system in the traditional way and users' judgment of relevance. It was clear that aboutness or topicality is not a sufficient condition for relevance judgment, since many other variables interact in making a document relevant in a given situation (Harter, 1992). Indeed, users may have quite different needs even if they submit identical queries. The user's mental state and the cognitive effect of acquiring knowledge were increasingly stressed (Ingwersen, 1992; Kuhlthau, 1993). Relevance assessment seemed to be closely tied to the user's previous knowledge and experience, current cognitive state and perception of the situation that triggered the need, explaining why relevance cannot be predicted by others (Morris, 1994). This knowledge and cognitive state, and even the user's perception of the situation, necessarily change during the process, making the information need highly dynamic and fluctuating. Also, different users are able to extract different amounts of knowledge – and even different knowledge – from the same documents. For instance, if a text is written in a highly technical style it may not be very informative to a given reader even though a specialist might insist that it is highly relevant. Finally, documents that are easier or cheaper to access may be considered

[73] We shall have more to say about this in section 4.3; the debate has still not been settled as regards the possible existence of something that could be called objective relevance. We shall find that there are indeed certain kinds of relevance that are objective, but the notion is only useful in very restricted circumstances. Relevance is fundamentally subjective.

more relevant simply for this prosaic reason. There would seem to be no end to what may affect the relevance of a document!

In spite of the intense efforts invested, many questions remained unanswered, and as late as the 1990s there was still nothing like a consensus about the nature of relevance. Studies were carried out in which users were asked directly about their relevance criteria, and it is clear from these experiments that there is a considerable overlap between the criteria used in different contexts, which means that it should be possible to form a general, domain-independent theory (Barry, 1994). The last decade of the millennium thus brought a clearer understanding and overview of the findings of the previous decades, but no satisfactory solution to the deep questions that relevance researchers so passionately care about.

4.2.2 Relevance as a multidimensional phenomenon

In her review of the literature on relevance, Borlund (2003) emphasized that relevance is multi-dimensional as well as dynamic, a view that is by now generally accepted. However, while succeeding in illustrating the overwhelming complexity of relevance and the differences between the various understandings, she failed to explain what unites them into one complex and fundamental phenomenon. How can one phenomenon be about being on the right topic, being informative, being useful and being easily accessible all at once? In that particular regard, Saracevic fared somewhat better when he concluded his 2007 review of relevance research by summing up the five *manifestations* of relevance to which all the different relevance notions reported in the literature could be ascribed[74]. His manifestations represent the state-of-the-art understanding of relevance.

Saracevic (2007)

- **System or algorithmic relevance**: "Relation between a query and information or information objects in the file of a system as retrieved or as failed to be retrieved, by a given procedure or algorithm. Each system has ways and means by which given objects are represented, organized, and matched to a query. They encompass an assumption of relevance, in that the intent is to retrieve a set of objects that the system inferred (constructed) as being relevant to a query.

[74] He originally published these ideas in Saracevic (1996), but the overview below is quoted from the slightly amended version in Saracevic (2007).

Comparative effectiveness in inferring relevance is the criterion for system relevance."

It is not clear that this dimension is really relevance at all; its value depends entirely on the algorithms in the system rather than on any real relevance criterion. If something judged as relevant in this way is actually useful (which is hopefully the case sometimes), this is a secondary result of the system designers having considered other, real relevance criteria which relate to actual relevance.

- **Topical** or **subject relevance**: "Relation between the subject or topic expressed in a query and topic or subject covered by information or information objects (retrieved or in the systems file, or even in existence). It is assumed that both queries and objects can be identified as being about a topic or subject. Aboutness is the criterion by which topicality is inferred."

The term *subject relevance* should not be confounded with *subjective relevance*; it is in fact a kind of "objective relevance" as it involves matching of labels according to the principle of aboutness. It is a notion that originates with library practices.

- **Cognitive relevance** or **pertinence**: "Relation between the cognitive state of knowledge and [cognitive information need][75] of a user, and information or information objects (retrieved or in the systems file, or even in existence). Cognitive correspondence, informativeness, novelty, information quality, and the like are criteria by which cognitive relevance is inferred."

This is similar to the view advocated by discourse analysts adhering to Relevance Theory (Wilson & Sperber, 2004), and was probably inspired by it. It is also closely related to the views held in cognitive IR (Ingwersen, 1992).

- **Situational relevance** or **utility**: "Relation between the situation, task, or problem at hand, and information objects (retrieved or in the systems file, or even in existence). Usefulness in decision making, appropriateness of information in resolution of a problem, reduction of uncertainty, and the like are criteria by

[75] Through a mistake, these words were not included in the 2007 version and consequently have been reconstructed from Saracevic (1996) in order for the sentence to make sense.

which situational relevance is inferred. This may be extended to involve general social and cultural factors as well."

This manifestation derives from a tradition of studying the situational uses of information, one of the earliest goals of user-oriented IR (Harter, 1992).

- **Affective relevance**: "Relation between the intents, goals, emotions, and motivations of a user, and information (retrieved or in the systems file, or even in existence). Satisfaction, success, accomplishment, and the like are criteria for inferring motivational relevance. It can be argued that affective relevance underlies other relevance manifestations, particularly situational relevance."

This last manifestation is much more nebulous and less easy to grasp than the others, but it is undeniably there; its accept can be traced back to the beginning of the *sensemaking* paradigm (Kuhlthau, 1993).

The model proposed below also includes manifestations or *dimensions* such as these, but the actual choice of dimensions differs. Also, they are placed in relation to each other. Saracevic, too, had originally pointed out the importance of an increased understanding of the interplay between the manifestations.

> The effectiveness of IR depends on the effectiveness of the interplay and adaptation of various relevance manifestations, organized in a system of relevancies. Thus, the major direction of R&D in information science should be toward increasing the effectiveness of relevance interplays and interactions. This should be the whole point of relevance research in information science. (Saracevic, 1996, p. 216)

Even though he repeatedly stresses the interactive nature of his strata, the idea is only just touched upon in the last sentence of his overview, "It can be argued that affective relevance underlies other relevance manifestations, particularly situational relevance". Here, we shall elaborate and extend the principle to all dimensions of relevance so that each dimension can be seen to precede or "underlie" others, except what amounts to Saracevic's *affective relevance*, which in the somewhat broader form advocated here constitutes the last word in the assessment process and as such rather overlies the others than underlies them.

4.2.3 Relevance as a stratified phenomenon

Based on the philosophical work of Alfred Schütz, Saracevic (1997b) suggests that a stratified model of relevance be constructed. He goes on to propose one which has been rightly acclaimed as the most satisfactory model to date. In his words, for a model to be stratified means "that the object modeled is considered in terms of a set of interdependent, interacting layers; it is decomposed and composed back in terms of layers or strata" (Saracevic, 2007, p. 1926). The manifestations of relevance illustrate this stratification. He proposes the elements in fig. 4.2 as a first attempt at capturing the strata. Other authors often refer to this as *Saracevic's stratified model* as if it were an actual model, and indeed this is the way it was presented in the original proposal from 1996, but in later works Saracevic himself seems to be quite clear about its status as no more than an initial sketch.

> The figure is a graphic depiction of the model, to be considered as an illustration of elements, variables, and processes involved, rather than an inclusive enumeration and specific ordering. The strata are not necessarily imbedded within each other, nor do they form a hierarchy. The relations between strata are much more complex and **could be in flux**. The user side has a number of levels. I suggest three to start with: Cognitive, Affective, and Situational. The suggested computer levels are Engineering (hardware), Processing (software, algorithms), and Content (information resources). It should be recognized that each level can be further delineated or that others may be added, depending on the given set of conditions or emphasis in analysis. (Saracevic, 2007, p. 1926; emphasis added)

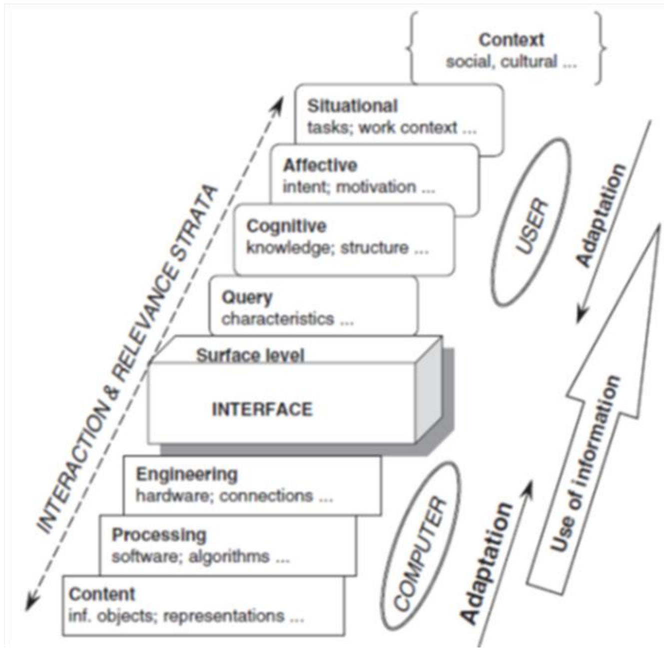


Fig. 4.2. Saracevic's stratified model.

The figure indeed seems to be an unfinished attempt at a model of the search process with all that this involves, rather than a focused model of relevance, and this idea is confirmed by the title of the follow-up paper from 1997, in which he calls it *the stratified model of information retrieval interaction*. Nevertheless, it is certainly an improvement over earlier attempts at providing a model of relevance, and the one developed here is based on stratification principles that echo those advanced by Saracevic, though the actual elements differ.

The **Regulated Flux model** of relevance proposed below comprises six interdependent dimensions of which four resemble the last four manifestations mentioned by Saracevic (fig. 4.3).

Suggested dimensions	Saracevic's manifestations
Desirability	**Motivational** or **affective** relevance
Utility	**Situational** relevance or **utility**
Topicality	**Topical** or **subject** relevance
Informativity	**Cognitive** relevance or **pertinence**
Practicality	none (not considered, or possibly as part of pertinence or utility)
Intentionality	none (not considered)
none (considered a mere proxy for relevance)	**System** or **algorithmic** relevance

Fig. 4.3. Rough correspondences between the state-of-the-art concepts as typified by Saracevic's (1997b) *relevance manifestations* and the *dimensions* suggested here.

We shall prefer the term *dimension* to *manifestation* because it emphasizes the fact that all dimensions are (or at least may be) present at once and act in concert, as indeed seems to have been Saracevic's intention[76]. In the suggested model, the interplay among dimensions is clarified by identifying at which stage in the search process each is active and available for reassessment. The dimensions adopted here are **informativity** (similar to cognitive relevance or pertinence), **topicality** (topical relevance), **utility** (situational relevance or utility) and **desirability** (related to affectiveness). To this is furthermore added **practicality**, which seems to be a necessary, though rather prosaic, ingredient. It covers resource characteristics such as access price, readability of electronic format, geographical distance that it is necessary to travel in order to access the resource, etc. A final dimension to be added is **intentionality**, which is more abstract and requires explicit explanation. The dimensions are introduced properly in chapter 6.

4.2.4 Relevance as a relation

Relevance has long been recognized to be basically a relation between two things, and the basic distinction between algorithmic and non-algorithmic relevance has incurred a tradition of considering

[76] The term *dimension* seems to originate with Schamber *et al.* (1990).

different kinds of relevance to be fundamentally characterized by what is relevant to what. The opinion of what these entities are has changed through the history of the concept. In the systems view it was treated as a relation between a query or a topic on the one side and a document on the other. Later, other versions appeared in which the query was replaced by the user or a user need, or something else.

Saracevic sums up a multitude of relevance theories proposed through the years by various authors in the shape of the following condensed formula.

> Relevance is the A of a B existing between a C and a D as determined by an E,
>
> where A may be 'measure, degree, estimate…;' B may be 'correspondence, utility, fit, …;' C may be 'document, information provided, fact…;' D may be 'query, request, information requirement…;' and E may be 'user, judge, information specialist. (Saracevic, 2007, p. 1919)

Relevance is the *gage of relevance* (A) of an *aspect of relevance* (B) existing between an *object judged* (C) and a *frame of reference* (D) as judged by an *assessor* (E).				
A	B	C	D	E
Measure	Utility	Document	Question	Requester
Degree	Matching	Document representation	Question representation	Intermediary
Extent	Informativeness	Reference	Research stage	Expert
Judgment	Satisfaction	Textual form	Information need	User
Estimate	Appropriateness	Information provided	Information used	Person
Appraisal	Usefulness	Fact	Point of view	Judge
Relation	Correspondence	Article	Request	Information specialist

Fig. 4.4. A five-dimensional matrix meant to represent all extant theories of relevance. From Schamber, Eisenberg and Nilan (1990), based on Saracevic (1970; 1975).

This condensation seems in turn to be based on the definition of relevance provided by Rees (1966). A complete overview of the idea as published in Saracevic's early papers (1970; 1975) is provided by Schamber, Eisenberg and Nilan (1990) and can be represented schematically as in fig. 4.4. Based on this same assumption, Stefano Mizzaro (1998) introduced a model consisting of an entirely different set of dimensions both from what Saracevic had proposed and from what shall be suggested below. He recommended the following dimensions.

Mizzaro (1998)

- **Information sources**: The (substantial) *document, surrogate* (consisting of title, keywords, author, name, bibliographic data, abstract, etc.) or *information* whose relevance is judged.

These are the resources whose relevance is assessed, i.e. the C column in fig. 4.4.

- **Representation of the user's problem**: The query, or the real or perceived need, or any other need representation, depending on the model of needs adopted.

These are the entities that signal what is needed, and against which the resources should be compared in order to determine their relevance, i.e. the D column.

- **Time**: This dimension captures the fact that relevance assessments change over time.

- **Components**: Topic, task or context according to which the information source is relevant to the query or information need.

Mizzaro acknowledges that the fourth dimension is "a bit more complex, since the previous three sets were totally ordered, while this fourth set is only partially ordered", i.e. he actually feels that his model forces order onto phenomena that do not really obey the proposed rules. Mizzaro's dimensions sum up some of the factors of relevance which are mentioned time and again as crucial, but the model does not amount to a truthful description of the relevance mechanism itself. His *information sources* correspond partly to what we have called the *facets* of a resource (see section 2.3), but while the document and the information it contains are both valid facets that a user may be searching for and find more or less relevant, a surrogate can hardly have this role; it is merely a medium through which the user judges the relevance of one of the other two. Hence, the seeming systematicity of the model is deceptive. Furthermore, the representation of a user's need is merely a proxy for the actual need, which is by definition what the user will in the end find useful. We shall accordingly not consider need representations to be part of the relevance mechanism. Time is also not properly a dimension of relevance since time does not change the relevance of a resource. Relevance assessments may indeed change over time, but not *because of* it; only as a result of other conditions that change over time. In the model advocated below, time is included simply by making the model dynamic and allowing reassessment of some dimensions but not others, depending on the stage reached. Among Mizzaro's so-called *components*, the present work does not consider *task* and *context* parts of the relevance mechanism as such, but the significance of all three components is captured by various elements in the model. Mizzaro makes a valiant attempt at disentangling the constituent phenomena of relevance, but he does not succeed in producing a convincing model, due to his letting the strait-jacket of traditional understandings of relevance guide him.

Saracevic's summation of manifestations shows clearly that he, too, adopts the view that the different manifestations of relevance are characterized by what they relate to what: *Algorithmic relevance* relates a document term to a query term, *topical relevance* relates a document topic to a query topic, etc. He even does his best to cement the relation as the main distinction between types. In the present thesis, it will be argued that the dimensions of relevance differ in quality rather than relations; it is certainly true that topics are involved in topicality, that cognition is involved in informativity (or in Saracevic's terms, cognitive relevance/pertinence), and so on, but this is a description rather than a definition of the dimensions.

We shall take the view that relevance is always a relation between the contents of a resource (with their intentionality, practicality, informativity, topicality, utility and desirability) on the one hand and a user (with his or her possibilities and requirements, cognitive state and demands, topical focus and demands, situation of use and needs, interests and preferences, etc.) on the other[77]. The difference is in what the user is attending to, and the query is indeed a crucial clue to this attentional focus. However, since queries are both terms and topics, we cannot actually know whether the user is focusing on one or the other and which of them the query was meant to represent. It may even be something entirely different, such as a clue about the contextual situation in which the information is to be applied. Thus, by way of an example, the query *catalogue librarians* could mean several things.

4.1	**catalogue librarians**	*

[77] In terms of the definitions given in section 1.1.3 it is rather a relation between a resource and a context (including the user who defines that context).

It might be a request for documents containing the string in question; this could be the case if the user wanted to refind a page that he or she knew existed (or a text known from an off-web situation which might be found online as well), if the user wanted to determine the correctness or frequency of occurrence of the phrase, or simply in order to retrieve any pages containing that exact phrase in the hope that they would be helpful. Alternatively, it could be interpreted as the intersection of the topics CATALOGUE and LIBRARIANS, which could mean either catalogues *of* or *for* librarians. This would be a more situationally relevant type of search, depending more heavily on utility. Finally, it might be that *librarians* refers not to an external class, but *egophorically* – to borrow a term from Tibetan linguistics (Tournadre & Dorje, 2005) – to the user's own class, i.e. *I AM LOOKING FOR A CATALOGUE SUITABLE FOR A LIBRARIAN SUCH AS MYSELF*[78]. Which type is present in a given case is hard to know unless one is prepared to ask the user explicitly, but at least we can try to understand what the different requests mean, as a message conveyed from the user to the system. Extracting the actual parameters is the second step in such a process.

4.3 A subjective nature

The past sixty years have seen an intense debate as to the character of relevance, the extent to which it is subjective and the ways in which this subjectivity manifests itself. The following sections elaborate on some of these characteristics. Relevance is both subjective, dynamic and a matter of degrees. These are important considerations before we set out to develop a new model of reference incorporating these aspects.

Relevance is dynamic in that it changes during the process of acquiring information as the user's mental state undergoes cognitive changes (Harter, 1992; Borlund, 2003). The same thing happens in a more typical discourse situation (Fauconnier, 1997). Studies of people in the process of writing research proposals confirm that different kinds of information are sought at different stages (Vakkari & Hakala, 2000). The significance of this finding becomes clear when you realize that every IR event may in fact be considered a miniature research project; it is always a process in which information needs develop dynamically. Consequently, the same behaviour is likely to apply in search generally. One such effect is that people tend to search for documents on a general topic at first, until they gain a more wholesome understanding of their information needs. After this stage is reached, however, they prefer documents that are relevant to the now more clearly perceived information need (Kuhlthau, 1993). In addition, partially relevant documents seem to be of relatively greater importance early in the process (Spink,

[78] We shall see in Part III that the most interesting feature of this kind of queries is actually not that they refer to the user, but rather that they refer to (the user in) the initial stage of the search, and not the desired end state.

Greisdorf, & Bateman, 1998). Thus, users' needs and preferences shift through the search process, and it is conceivable that search engines might be tuned to fit this development by adjusting to the stage in the process that a user has reached.

One reason that people's information needs cannot be objectively equated with a set of documents is that they are dependent on previously acquired knowledge. Even though people do sometimes look for documents that they have seen before (Harter, 1992; Dumais, et al., 2003), in most cases what is already known would intuitively seem to be irrelevant. This fact is well known in e-learning studies; once a lesson has been gone through, it is not actively presented again (Brusilovsky, 2003). However, as pointed out by Kuhlthau (1993), redundancy may in fact contribute significantly to the value of resources, and hence it should not be taken for granted that known resources are unwanted noise.

4.3.1 Objective relevance

In spite of the multitude of studies confirming that relevance is fundamentally subjective, it remains a standing discussion whether there is – in addition – something which might reasonably be called *objective relevance*. What makes this discussion especially important is the fact that the distinction between objective or algorithmic relevance on the one hand and the subjective types on the other is generally considered so basic that it is taken as a model for what relevance dimensions should look like, and this, the present author feels, has hampered the research of relevance severely, as the distinction is of an entirely different kind from all other relevance types.

As its relevance is judged by a human observer, a document cannot be truly objectively relevant. On the other hand, the IR system's act of suggesting it in response to a query may indeed be said to be objectively relevant as the relevance of that act is not dependent on an external observer but is based solely on the information available to the system at the time of the act. In other cases, objective relevance can be achieved artificially by excluding subjective data and basing the judgment on agreed-on parameters only, but this rarely results in a useful notion.

Despite the fact that each way of automatically assessing objective relevance produces a slightly different set of relevant documents (Borlund & Ingwersen, 1997), the objective view of relevance has always had its followers, and it remains tempting to fall back on such a view as it simplifies things enormously. Indeed, it seems a natural supposition that there should exist objective types of relevance. After all, that is how we typically think and talk of it; we say that a document is relevant to a certain topic or query, and we feel that this is indeed the case (Swanson, 1986). However, relevance is certainly subjective in the sense that it depends on the user's situation and view of the world. It may even be subjective to the extent that we can disagree on the relevance of a certain resource. To Swanson, this very fact indicated that there must be such a thing as objective relevance. If the phenomenon were truly subjective, it seemed to him that this ought not to be possible because that would make the requester's judgment correct by definition. However, it is not the case that "anyone who disagrees with the requester's judgment is necessarily mistaken" as regards the relevance of a stimulus (Swanson, 1986, p. 391). Rather, if both people involved had a *complete* (or at least identical) knowledge of the other's situation and view of the world, they would have to agree on the relevance of a given stimulus; it is just that it is not possible to obtain such complete or identical knowledge.

Swanson's concerns are easily placated. As part of our engagement in discourse we must assume or pretend that we share our interlocutor's point of view; this is a necessary precondition for communication to take place (Maturana, 1980; Alfort, 2009). We furthermore assume that we would find the same resources relevant as our addressee, and hence may argue on the relevance of a certain resource in what we believe is the same situation perceived objectively by two people. Of course, it so happens that we never share the exact same context, and consequently there may well be a mismatch between what two interlocutors find relevant, leading to discussions between them (Gadamar, 1960). This is not a sign that relevance is really objective, but merely mirrors the fact that we have to believe in the dream of perfect communication with a shared context and perspective, because we would simply not be able to communicate or otherwise interact in the absence of such an assumption. The objectivity of relevance is thus a hallucination of discourse and not a feature of relevance itself; relevance is no more objectively definable than truth or the true colour of an object (Travis, 1981).

> How two interlocutors accommodate their divergent perspectives, and negotiate the adaptation of conventional patterns to the idiosyncratic complexity of the immediate context, are foundational questions of linguistic theory. (Langacker, 2000, p. 389)

Swanson points to the Austro-British philosopher Karl Popper's notion of *worlds* as another indication that objective relevance does exist[79]. Karl Popper, who is renowned as the greatest philosopher of science in the twentieth century, distinguished between three different *worlds* (Popper, 1972; 1978): *World 1* is the physical world which is out there (whatever that implies) and can be perceived; *World 2* is the mental world representing our perception of World 1, including our ideas, beliefs and needs; and *World 3* contains products of the human mind such as books and works of art. These often inhabit World 1 as physical entities, but they can only be interpreted with reference to the human mind. Unlike thoughts and beliefs which can only be described, we can share and discuss members of World 3.

[79] Brookes (1977; 1980) seems to have been the first to have drawn the attention of the information science community to Karl Popper's work.

Swanson sees this as an indication that whereas information needs belong in World 2, queries are objectified needs that inhabit World 3; i.e. they can be shared, compared and discussed. The idea is good, but unfortunately it does not work as an argument against denying the existence of objective relevance because words are so much more than just strings of letters; a word contains much more information about the user's intentions than just a topic label. Hence, as is the case with books and works of art, queries can only be interpreted with reference to the human mind. If computers understood as much by a word as a human interlocutor does, it wouldn't be a problem, but they do not. The fact that we can share and discuss something does not make it objective and its existence an irrefutable truth; quite on the contrary.

4.3.2 Logical relevance

Several advocates of the idea that there are indeed phenomena out there that could be called objective relevance have tried to formalize the notion with a view to establishing a mathematically precise definition of what evidence is relevant to a question. These researchers have come from machine learning, knowledge representation and artificial intelligence, philosophy, economics, and even forensics. To Cooper (1971), whose ambitions lay in designing inferential systems, a sentence (i.e. a piece of information) is *logically relevant* to the user's information need if and only if it contributes a node to a tree of pieces of evidence from which the answer to the user's question can be inferred. In machine learning, Bell (2000) refers to Gärdenfors (1978) for whom any evidence which affects the likelihood of a given hypothesis being correct is relevant to that hypothesis. This is not as different from Cooper's definition as it might seem, since from the moment the user has consulted the system, Cooper, too, deals with sets of potential answers (i.e. hypotheses) rather than actual questions. Finally, in the legal sciences, a piece of evidence is relevant to a case if it contributes a node to a diagram of premises from which the answer to the question of guilt can be inferred (Callen, 2003 Mich. St. L. Rev. 1243). Again, this is basically the same thing applied to a different domain.

From the point of view of (non-inferential) IR, these definitions of logical relevance fail to be useful. The reason that objective relevance rules work in forensics or machine learning is that the arguments are based on formally stated premises which explicitly establish the worldview on the background of which the relevance of a given piece of information is evaluated. This is not the case in communication, and hence not in IR. The above definitions all depend on the basic assumption that what the user wants is an answer to a question or an evaluation of a hypothesis, and current IR systems do not build on this assumption but rather let users search for any kind of information even when they cannot specify a particular need. In such cases, logical relevance is simply not a useful notion. Cooper (1971) is of the opinion that logical relevance can be extended to cover topical queries insofar as these are reinterpreted as representing the totality of possible questions about the topic. However, he significantly hedges this suggestion by adding "provided the topic query reflects a clear-cut information need on the part of the user" (Cooper, 1971, p. 32), clear-cut being exactly what information needs rarely are. Consequently, a different kind of relevance is needed when dealing with IR.

4.3.3 A relevant act

LATs are computer systems and consequently cannot reason beyond what they are told, and thus objective relevance does make sense in at least one case. Even if we are not satisfied with the results presented to us by a search engine we do not necessarily think that they were bad suggestions (i.e., that they were irrelevant results to give), but only that we did not find anything among the suggested resources that was *really* relevant to us. That is, the act of suggesting the document may be relevant even though the actual document is not helpful in the user's situation, since the act of suggesting a resource does not refer (directly) to the user but to the information available to the system. This is one of the rare instances where objective relevance makes sense: Relevance can be objective if it is based on the exact same premises in all cases of evaluation, and in this instance this is true, because we explicitly avoid referring to subjective data not accessible to the system when evaluating the performance of the system and concluding that its actions were relevant. This difference in relevance between a piece of information and the act of giving it may also be observed in certain kinds of evaluative discourse.

4.2	A: You did not write to me! B: That is not relevant at present. A: Excuse me! I rather think it is.

What the speakers disagree on in such exchanges are the underlying inferences; they value criteria differently and therefore disagree on the relevance of one criterion. In other words, B does not really object to the relevance of A's comment, but rather to the relevance of the underlying question of whether B wrote to A. B does not value letter-writing as highly as does A, and as such differences are a matter of taste and opinion they cannot be easily parameterized. The two interlocutors would have to agree, however, that on the premise held by A that writing was important, A's comment itself was certainly relevant. Thus, the relevance of an act of communication may differ from the relevance of the actual information conveyed.

If the same document were given to the same person twice, we might expect the relevance of the document to be the same in the two instances. On the other hand, the premises are of course *not* identical – the second time the reader will already know what information the document contains, and hence the relevance of reading it again will be significantly lowered. Thus, as soon as people are involved, the process is considerably more complex and subjective. In everyday life we act as if we share premises, because this is a prerequisite for communication to be possible, while in actual fact we never do share the exact same information, views and perspectives and consequently objective relevance is only a useful notion if all differences in perspective are explicitly levelled so that it becomes a question of relevance by definition. Thus, in a library context a textbook on poetics may be considered to belong to the topic of poetics by definition, even though individual librarians may disagree on the appropriateness of filing that particular book under the label in question. This is again a case of objective relevance, and it is very much on the borderline of being an instance of relevance at all. It is not very helpful to have a notion of relevance which implies that when you ask for a book on poetics, any book filed under that topic is considered relevant by definition, without even considering what it is really about. However, your request for the book will make the librarian go to the section and fetch the book for you, and this act of hers may indeed be considered relevant, given that the book was in the *poetics* section and hence *mutatis mutandis* ought to be relevant to your request. Thus, the act of suggesting a resource based on an objective (definitive) classification is a case of objective relevance that actually makes sense. To what extent it is helpful is a different matter.

Similarly, when a LAT returns a result, this act can be said to be relevant, given certain premises; compare the relevance judgment *Was this document useful to you? – Yes, because it told me about the topic I was interested in* with *Was this suggestion useful to you? – Yes, because it pointed me to a document which the system thought might be useful to me, and that was very helpful since I have had difficulties finding useful information*. The system suggests a resource based on the information available to it, and if that points to a certain document as being a likely candidate for a relevant resource, suggesting it is an objectively relevant action. The user may disagree on the premises, of course, but he or she will have to agree that given the information available to the system, the response *was* relevant. Thus, if the system suggests a document which the user finds irrelevant, where the user disagrees with the system is in the underlying inferences built into the algorithm or the correctness of the data available to the system rather than in the relevance of the act of suggesting the document, and it is important to be clear about what we are talking about when discussing relevance evaluation.

Given that computers would be expected to *always* perform objectively relevant acts, as long as there are no errors in the script, in order to be a useful notion objective relevance must refer to the other kind – the relevance that is established by definition, typically between topic labels and text on those topics. However, definitions vary between cultures, so relevance is still only objective to the extent that we implicitly assume that we share the same cultural premises, which is by no means certain. For example, a biologist may think that documents that mention *bananas* and *tomatoes* should be objectively relevant to the query *berries*, given that biological taxonomies classify them under that label, but others may disagree; the owner (or user) of a pet shop may think that *rats* are objectively relevant as a subordinate concept of *pets*, whereas others might prefer to group the rodents in question under *pests*, and so on. Thus, objective relevance evaporates as soon as the shared premises are revealed not to hold, and this is likely to be a situation encountered much more frequently than we are usually aware.

An IR system cannot possibly add anything beyond the information given to it, and so its decisions represent the absolute truth as far as it is concerned. Note, however, that the system-driven approach to information retrieval which is based on objective relevance does not necessarily ignore the user's specific needs; insofar as the information need as perceived by the user and the task he or she is facing can be parameterized, they too can be incorporated into the algorithm through a user profile. If it were possible to identify all parameters which could possibly affect a human's relevance judgments, other kinds of relevance might likewise be considered objective when looked at from outside the judging person, because all the necessary information would be present and the outcome would hence be certain. Thus, the feeling that a certain document is "objectively relevant to that topic" might be extended to "this document is objectively relevant to this kind of situation", and so on. However, that scenario is not achievable in the immediate future, and consequently all kinds of relevance which involve humans must remain subjective, for now at least.

Algorithmic relevance (relevance to a query) sometimes comes close to simulating the relevance of a resource, and it is extremely simple compared to the real thing. However, given that premises are never identical across evaluating subjects, algorithmic relevance and relevance to a topic or an objectified need with no explicit reference to the user are here treated as mere proxies for the relevance of a resource rather than the relevance itself. Proxies can be very useful, in fact vital, since measuring real relevance is highly problematic, but it is important to be clear about when it is justifiable to talk of real relevance.

4.4 The transfer of information

Just as the act of suggesting a resource may evince a different level of relevance from the actual resource itself, so giving information, or *informing*, may in fact take place even when there is nobody who is actually being informed by the information. This is an aspect of information that Buckland (1991b) takes up. He introduces a simple distinction between **information-as-process**, **information-as-knowledge** and **information-as-thing** which deals with the various everyday uses of the term without going into metaphysical discussions of the actual nature of information as such. What he calls *information-as-thing* is what we have called a substantial or formal resource (cf. section 2.3) containing *information-as-knowledge*, which in turn relates to the representational and interpersonal resource. *Information-as-process* is closely associated with the interpersonal resource which takes the sender and recipient of the information into account. On the other hand, an author has simultaneously a (subconscious) intention to *affect* via the effective resource, to *inform* via the interpersonal resource, to *portray* via the representational resource, to *write* the formal resource, to *produce* the substantial resource and to *reach* an audience through the intentional resource, and a full understanding of a document requires the reader to decode and acknowledge all of these intentions. As pointed out by the French cognitivist Dan Sperber and the British psychologist Deirdre Wilson who founded Relevance Theory, communication not only involves conveying a message but also conveying the intention of conveying the message, so information without the coding and decoding (interpretation) of the author's intention to inform does not constitute successful *communication*. This amounts to what Wilson and Sperber call the *informative* and the *communicative intention*, respectively; "understanding is achieved when the communicative intention is fulfilled – that is, when the audience recognises the informative intention" (Wilson & Sperber, 2004, p. 255). Without the communicative intention being fulfilled the reader may still be able to derive information from the resource, but only incidentally.

4.4.1 The nature of information

The nature of *information* has long been a subject of debate. It is certainly a many-faceted notion, and the definition that one chooses to adopt must reflect the uses to which one intends to put it (Machlup, 1983; Ingwersen, 1992). We shall therefore have to decide on a definition before moving on to the discussion of topics. The controversies mainly concern the distinction between data and information, i.e. whether information proper requires an intending sender and/or an interpreting recipient, but also whether it must affect the recipient's understanding in a positive way in order to qualify as being really information.

In the prototypical case, information involves an intending sender communicating the information in question to an interpreting recipient (Machlup, 1983; Brier, 1996). The presence of a sender makes a crucial difference because a sender always provides a selection of pertinent data to be used in a certain context rather than merely supplying random data to be interpreted at will.

> Documents often contain data, of course, but it is data that has been selected, interpreted and presented *for use* – it is data *in context*. (Blair, 2006, p. 78; emphasis original)

Whether the recipient is able to interpret this data and make use of it is quite another matter; one that is highly relevant when the recipient is a computer. Ingwersen (1992) argues that for the purposes of IR research an intending sender should be taken to be a requirement for information to be present. I would like to reformulate this as a requirement that the sender be included in the conceptualization of information, without necessarily being present in all actual instances.

As I see it, the prototypical scenario gives rise to two intersecting notions of information. One is that it is something originating with an intending sender and intended to be interpreted in a particular way, and the other that it involves data which through interpretation are able to change the recipient's understanding of something. Ingwersen and Järvelin (2005a) are of the opinion that both of these criteria are required for something to be considered information, and while the various schools tend to focus on either side, we shall take both to be equally important ingredients of the phenomenon. However, to me it seems that either of them is enough. The consequence of this encompassing view of information is that it does not necessarily have to involve both a sender and an interpreter in practice, but it must have either, and the conceptualization of it must include both.

Information involves an intending sender and/or a recipient who is cognitively affected by it.

To me at least it does not seem wrong to talk of information as arising from the unilateral interpretation of data – such uses simply represent a natural extension of the prototypical sense of the word. If there is information, there is also a sender, at least *in principle*, but the sender may be an abstract one that is not really there, i.e. in the right kind of context, data may take on the role of a sender and speak for itself, not only metaphorically, but in a very real sense. Thus, when an antelope senses the presence of a lion in the grass, this is not information that is willingly sent by the predator, but the antelope is nevertheless able to interpret the data provided by its sense organs (*for use in* *context*) and derive useful information from it. This kind of information is just as important as the one that is truly communicated from an intending sender to an intended recipient. In principle, one might suggest that the sender in this case was the sense organs and the subconscious, but such an explanation would be applicable in all cases of potential information, which would make the concept of *sender* vacuous – there would never be any need to recognize a sender, and consequently a recipient would always be enough. A sender is thus not strictly necessary, as long as the data connect with context and make sense (i.e. produce a *positive cognitive effect*, in the parlance of Relevance Theory) in that context. Thus, information might be said to be *contextualized data*.

The opposite situation where there is a sender but no interpreting recipient is somewhat harder to defend. Indeed, it is common in the cognitive tradition of IR research to deny the existence of this kind of information (Brier, 1996).

> (...) messages sent by machines or man to a human being can become information in the real sense; however, signs communicated by machine or man to machines can never become information, although they are perceived and affect the embedded cognitive structures. Such signs stay signs - or remain as 'potential information' - at a given linguistic surface level. (Ingwersen, 1996, p. 7)

Since a human sender is automatically also an interpreter of his or her own information[80], a decisive case would have to be sought among messages given by computers to human recipients but not actually received by anyone. Imagine an information system able to extract information from a large corpus and deliver it in an automatically compiled e-mail to the requester. If someone were to ask for *everything ever written on the topic of computers*, and an answering e-mail was delivered to that person a week later but for some reason was never opened, we would presumably not like to say that it did not contain any information just because nobody actually read it. Thus, like the antelope, a machine can generate *information* about something from gathered data, even though it is not a prototypical interpreting recipient. The case should be compared to the discussion of objective relevance above; just as a stimulus cannot be said to be truly objectively relevant but the act of suggesting it can, information cannot be said to be really informative unless interpreted by a subject (i.e., it remains *potential* information), but you may still talk of *giving* information even when the interpreter is absent.

[80] Note that this observation complements the one that a human interpreter (in the shape of sense organs and nervous system) is automatically also a sender of the information. This circularity makes information a cybernetic phenomenon.

It is reasonable in this connection to speculate whether information has to be relevant for it to be real information. Specifically, does something that you already know constitute information? In principle, the change from data to information happens when it is received by a cognitive system and makes a difference to the state of that system.

> Data first become information when they are integrated with a given knowledge process and pre-understanding. That means that they only become information when they are interpreted in a bio-psychological-social knowledge system. (Brier, 1996, p. 302)

In keeping with this, second-order cybernetics researchers define information as "'a difference which makes a difference' [to] a living autopoietic (self-organised, self-creating) system" (Brier, 1996, p. 296)[81]. In our view, the recipient does not have to be *living* in the sense normally ascribed to that word, but can be a computer or other cognizing recipient in a wide sense. This is another extension of the prototypical sense of the term because when there is information it should in principle change the recipient's understanding of something, and it is not clear that a computer can really be said to possess understanding[82]. However, it is certainly still able to make inferences and act on the basis of the information in question, and so it may be considered to have an abstract kind of understanding that has been affected in a way roughly similar to that of a human recipient. One should remember that we are testing the limits of the concept here, while in most cases it is less difficult to decide whether something is information.

A compromising answer to the question of the necessity of real informativity in the spirit of the definition that we are advocating here would be that all data is also information to the extent that it is seen as having the *potential* to inform, i.e. to change someone's understanding of something. This is true of the *unread e-mail* example as well as of the antelope's interpretation of sensory data, and also of more prototypical cases of information transfer.

Information is data that has the potential to inform, either because a sender has the intention to inform, or because a recipient is cognitively affected by it.

[81] The idea of information as a difference that makes a difference originates with Bateson (1973).
[82] As mentioned earlier, we shall say that it does indeed possess understanding, though of a much simpler type than beings with brains, cf. Section 4.6.3 on levels of complexity in cognition.

It follows from this stance that something is only information *to someone*, and that even though it is not (relevant/new) information to you, as soon as it is considered as being interpreted by someone else, it may again become information, because now it is informative. Thus, information does not have to make a difference, but it must have the potential to do so. In this light, the reason that information may at times exist with no actual recipient is that it is still *potentially* informative and thus effectively qualifies as information. In its essentials, therefore, one might say that information is indeed in a sense dependent on an interpreting recipient, though this recipient need not actually be present.

It would now be prudent to attempt to apply this result to the case of relevance: If "objective" information is really just potential information, then it follows that *objective relevance* might alternatively simply be **potential relevance**, and indeed such a stance would seem appropriate.

4.4.2 Informativity

While *information-as-thing* is a concrete manifestation of potential information, *information-as-process* specifically involves changing the recipient's understanding of a topic. This is of course a matter of degree, and it is also a question of benefits as opposed to simply creating more confusion by adding information. In other words, information can be more or less **informative** in various ways. The informativity of a document is affected by the amount of information it contains, as well as the scope and depth of it.

> [A] user needs the data to be broad enough to satisfy all the intended uses, and at the same time to include no unnecessary information". (Xu & Chen, 2006, p. 963)

Furthermore, the novelty of a document from the point of view of the user is important, as known information is, *ceteris paribus*, considerably less informative than new information. In fact, as noted above, one might well ask whether information that the user already has acquired (as information-as-knowledge) is really information (to him or her), seeing as it is not really informative. The answer must be that a document containing such information is still information-as-thing with the potential to inform, but (assuming that the information was properly incorporated in the reader's knowledge stock at the first perusal) it no longer gives rise to information-as-process for that person, which is equivalent to a lack of informativity, whatever the other benefits of redundant information. In sum, then, users are primarily looking for information-as-thing that is likely to trigger information-as-process leading to an expansion of their information-as-knowledge. It is the task of a search engine to help users in this endeavour.

> Search engines should help users find information-as-thing that is likely to trigger information-as-process leading to an expansion of their information-as-knowledge.

Also – and this is not usually mentioned in accounts of relevance in IR, though it features prominently in discourse studies – some information may cause more vital changes to the user's understanding of the topic than others (Wilson & Sperber, 2004). If crucial inferences can be made, the information is more informative. For instance, if someone is looking for information on Denmark, a statement mentioning the Danish prince consort is – taken in isolation – more informative than one referring to the queen, because while both may give rise to the inference that Denmark is a monarchy, only the former enables the reader to infer that the queen is the actual monarch.

Next, the user needs to be able to understand the contents of a resource in order for it to be successfully informative. This is a parameter that hinges on the user's prior knowledge, abilities, experience with the topic and the style in which the information is communicated, and also the clarity of the text; informativity is thus a highly complex and subjective dimension. Even emotional features such as trust are paramount; if the user does not have confidence in the reliability of the source the information is effectively not accepted, and this blocks informativity. Sometimes this same effect may arise from disagreement with or sheer dislike of the author, or an antipathy towards the topic leading to avoidance of a resource and consequently blocking informativity. Even the expressive quality and typographic layout of a document may influence informativity since a boring text or unappealing document (technically, an undesirable intentional, substantial, formal, representational, interpersonal or effective resource) may deter the user from reading and understanding the contents properly.

It is clear that many different factors interact in making a resource more or less informative. The following list mentions some of them.

- Amount of information
- Scope and depth of information
- Inferential importance of information
- Novelty of information to user
- Ability of the user to understand the contents
- Visual and verbal clarity

- User's trust in source and contents
- User's accept of information

All of these are characteristics of the information, the document or the user that add to the informativity of a resource. Contrary to existing accounts of relevance, we shall not say that they affect the relevance of a resource directly, but rather that informativity in turn plays a crucial role in relevance. This will allow us to generalize more effectively and include any unforeseen criteria of informativity under the umbrella of relevance in advance. Only too often has the discovery of new criteria and behaviour patterns led to unjustified extensions of a model in different directions in order to accommodate them, resulting in a proliferation of random sprouts and patches[83]. For instance, it is likely that background and text colours play an important role in readability of the substantial document. If it should turn out that certain colours are particularly inducive to reading for psychological reasons, this would not change the entire understanding of relevance, because it would simply be an element contributing to informativity, which is already part of the model of relevance as developed below.

4.4.3 Cognitive gains and degrees of relevance

Ever since Sperber and Wilson formulated their Relevance Theory, degree of relevance has frequently been connected to the amount of changes to someone's understanding that a certain resource is able to affect as well as the efforts required in order to extract the information. Reading information demands effort on the part of the user, and so it is necessary to weigh the effort required against the probable cognitive gains. If the gain is likely to be small, users will not bother with reading the document unless this involves very little effort. On the other hand, if it seems to be just what they need, they may spend hours studying the one resource. Wilson and Sperber have investigated this effect as it occurs in ordinary discourse. Their basic idea as regards degrees of relevance is that these are the effect of varying amounts of cognitive gain vs. processing efforts (Wilson & Sperber, 2004). If the cognitive gain is substantial and the processing efforts are small, the stimulus is highly relevant; if, on the other hand, the gains are limited and not worth the effort then it is much less relevant. All other measures of degrees of relevance are in this view considered to be mere proxies for this, which may be useful as long as we are unable to measure the actual cognitive impact of information. We shall,

[83] One example of such a model is the one proposed by Mizzaro (1998), where highly different phenomena are forced into a constrictive model that treats them all alike. He recognizes four dimensions: Information resources, information needs, time, and the components topic, task and context. It is hardly feasible to treat all of these as being dimensions in the same sense.

however, be looking closer at different types of relevance grading in section 6.3, and we shall find that in IR at least there are several scales along which to measure the degree of relevance and these cannot be reduced to one except in the shape of their various contributions to the ultimate relevance evaluation where the relative importance of the many different criteria is applied and the resource is either accepted or rejected. Indeed, as soon as the surface is broken a single scale would not seem to suffice for a description of discourse either. Consequently, the results arrived at here can be applied to discourse analysis and enrich the understanding of relevance in that context by adding several scales of relevance gradation.

Apart from the efforts required, there are clearly degrees of changes to the reader's understanding induced by informative resources in the first place; seemingly quite unrelated information might potentially set off a process of increasing the understanding of different but related topics which could eventually lead to an improved understanding of the topic under scrutiny, but not to the same extent. Furthermore, as everything is connected you might say that every single document which contains information that the user can read is also relevant to some degree, given that it will to some extent bear on the topic in question or a related topic. For instance, as pointed out by Cooper (1971, p. 33) in connection with the query term example *halogen elements*, "virtually everything ever written is 'about' halogen elements in some sense or other. The U.S. Constitution, for example, is about halogen elements, among other reasons because it bears on the political problems of water fluoridation, a process which utilizes a halogen element". The majority of the available documents will have a relevance value very close to null, but only those resources which are entirely incomprehensible to the user or which are deemed irrelevant by definition due to certain user preferences are *completely* irrelevant. On the other hand, only those documents can be said to be completely relevant which are deemed to be so by the user. Many resources will of course have relevance values in between these two extremes (fig. 4.5). It is clear, for instance, that the U.S. constitution is exceedingly unlikely to be considered useful by the user who searched for *halogen elements*.

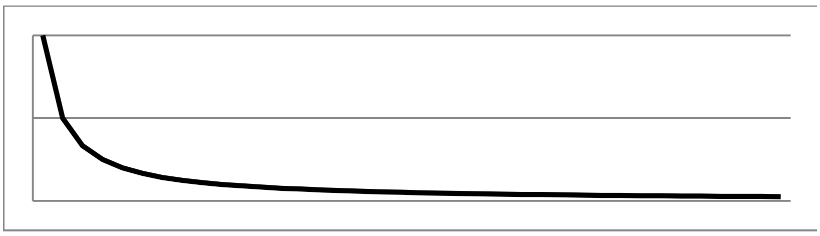


Fig. 4.5. Probable approximate distribution of *relevance* in some imaginary corpus, given a perfect way of measuring the degree of relevance of a document.

An alternative approach might be to measure the **irrelevance** of a resource (fig. 4.6). Zero irrelevance would then represent the perfectly relevant document. The distance from this point would signify a semantic distance from the central topic as well as the contribution of other relevance criteria to be studied below. The relative strength of various relevance dimensions and measures would have to be defined, which cannot but be highly controversial since people are sure to have diverging opinions on this.

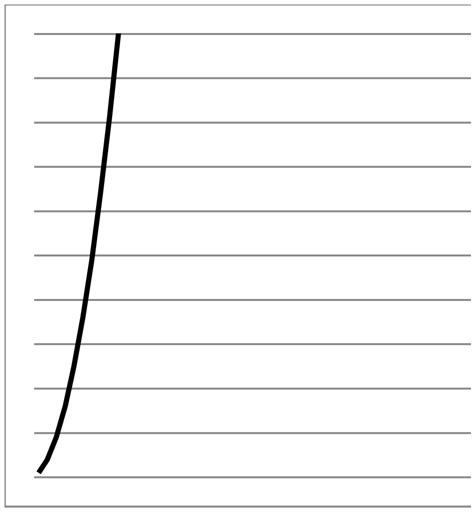


Fig. 4.6. Probable approximate distribution of *irrelevance* in some imaginary corpus.

It is hard to see how something could ever be perfectly relevant in the context of a topical search, unless this is so by definition as soon as the user judges a resource to be relevant. On the other hand, if a simple fact is sought, then one might well think that the provision of the needed information would make a certain resource perfectly relevant, but in fact it is merely *perfectly sufficient*; given that there will be features of the resource which could be graded along several scales, such as ease of access, attractiveness of layout and so on, it is unlikely that the imaginary ultimate measurement of irrelevance would ever reach a perfect zero, and this would be observable as an ability of users to grade various resources which all provide the same factual information but in different ways relative to each other as being more or less perfect answers to the need.

Traditionally, most systems focus on the basic dichotomy of *relevant* vs. *non-relevant* because this binary distinction makes it possible to measure the effectiveness of a given system using precision and recall (Kekäläinen & Järvelin, 2002). However, in empirical user-studies participants are nowadays often given the choice to rate a document as partially relevant, and new measures have been developed in order to make it possible to capture such cases. As already mentioned, it has even been pointed out that partially relevant documents may be crucial to users, especially in the early stages of the search process, and hence a system should not necessarily focus on retrieving fully relevant documents at the expense of the partially relevant ones (Spink, Greisdorf, & Bateman, 1998).

4.4.4 A maximum of relevance

Actually measuring the degree of relevance of a stimulus is of course highly problematic. Sperber and Wilson have formulated an equation in an attempt to visualize what this would entail in terms of their framework. Relevance, they say, is calculated as a positive function of the benefits the hearer would gain from processing the input, and a negative function of the processing effort this would require.

$$relevance\ of\ stimulus = \frac{cognitive\ gain}{processing\ effort}$$

Since they also claim that "cognitive efficiency in humans is primarily a matter of being able to select, from the environment on the one hand, and from memory on the other, information which it is worth bringing together for joint – and costly – attentional processing", i.e. which is relevant, this implies that humans must have some very effective cognitive mechanisms to calculate the relevance of an input (Sperber & Wilson, 2002, p. 14). The authors argue that, apart from the general ability to ignore irrelevant information in our everyday lives acquired through evolution, the human brain has evolved some specialised structures (though not necessarily actual centres in the brain) which are optimized for interpreting stimuli and identifying what makes something relevant.

> As a result of constant selection pressure towards increasing efficiency, the human cognitive system has developed in such a way that our perceptual mechanisms tend automatically to pick out potentially relevant stimuli, our memory retrieval mechanisms tend automatically to activate potentially relevant assumptions, and our inferential mechanisms tend spontaneously to process them in the most productive way. (Wilson & Sperber, 2004)

This does not mean that humans always succeed in establishing what is maximally relevant, but merely that we do on the whole manage our cognitive resources rather effectively as regards the task of calculating relevance.

As mentioned in Part I, Sperber and Wilson put forward the hypothesis that our thoughts are governed by what they call the *cognitive principle of relevance*.

Human cognition tends to be geared to the maximisation of relevance

Furthermore, they go on to formulate a *communicative principle of relevance*.

Every ostensive stimulus conveys a presumption of its own optimal relevance

For a stimulus to be *optimally relevant* it needs to be "relevant enough to be worth the audience's processing effort", and also to be "the most relevant one compatible with communicator's abilities and preferences" (Wilson & Sperber, 2004, p. 256). All utterances are *ostensive stimuli*, i.e. human language inherently conveys a presumption of relevance. Consequently, *irrelevance* presumably does not play any great role in communication, the only exception to this being the need to avoid listening to speakers who are talking to somebody else and whose utterances do not concern us in any way. This exception may in fact be much more important than is usually acknowledged, though communication studies tend to focus on laboratory-like exchanges with no discourse-external interference. In a realistic setting where people move around among each other and speak to each other, being able to avoid reacting to irrelevant stimuli is likely to always be of importance. Indeed, a series of studies performed in the 1950's show that if test subjects are presented with two different messages, one in each ear, and told to focus on only one of them, they are afterwards unable to recall the contents of the other message, or even the language in which it was spoken (Moray, 1959). The ability is particularly important in search, because here avoidance of the obviously irrelevant may involve going through pages and pages of suggested resources rather than a quick visual or auditory scan of the surroundings. This laborious filtering through what amounts to thousands of voices fighting for the user's attention can only be circumvented if we are able to teach the system to perform it for the user, to some extent at least, thus relieving the user of the task.

What we need is a theoretical framework which takes into account the subjectivity, dynamicity and multifaceted nature of relevance and unites it all into one coherent whole where all of these features are seen in relation to each other and make sense. First of all, however, we shall have to look closer at topicality and the nature of topics, first from a cognitive and subsequently from a linguistic point of view. We shall argue that irrelevance is paramount in all human processes of relevance evaluation, and that any IR system must take as point of departure the human way of experiencing our surroundings and processing the information gathered by our senses. This shall be the basis of the model of relevance developed in chapter 6.

4.5 Closing in on topicality

In spite of being foundational for all kinds of information technology, relevance is not a concept that has its roots in IT. Rather, it is a general *human condition* (Saracevic, 2007). In fact, the demarcation of the relevant may be said to be at the very basis of cognition itself.

> (…) relevance is a very basic human cognitive notion in frequent, if not even constant, use by our minds when interacting within and without in cases when there is a matter at hand. Relevance is a built-in mechanism, that came along with cognition. (Saracevic, 1996)

The notion has been studied widely, not only in information science, but also in fields as disparate as linguistics, communication studies, logics, epistemology, philosophy, law, economics, cognitive science, psychology, psychophysics and library studies, as well as e-learning, artificial intelligence and machine learning research. The concept has extremely wide applicability and pervades our interactions with the surrounding world; it holds together our speech and thoughts and tells us what to focus on in a world of infinite detail. It is also a concept that everybody uses constantly, quite effortlessly and without conscious thought. Even so, it continues to elude scientific definition.

Relevance has traditionally been thought intimately connected with – even equated with – *topicality* in the sense of *aboutness*, especially in the early phases of IR investigation. This turned out to be an unwarranted simplification.

> The idea that users want, or will be most happy with, documents on the topic of a search statement is an assertion that has absolutely no empirical support. (Harter, 1992, p. 603)

The question is what a topic really is; it is not immediately easy to come up with a definition of the concept[84]. Intuitively, it may be thought to be a word that sums up what a text is about. Yet it not only sums up the contents but also encompasses the abstract aboutness of a text. Consequently, a distinction can be made between, as it were, **topic-as-thing** (the word that sums up the topic, in IR the query term) and **topic-as-domain** (what a document is about), to remain in the spirit of Buckland (1991b). Strictly speaking, of course, there need not even be a document – topics (and queries) may occur on their own without necessarily matching any actual text. Nonetheless, they always sum up some potential information which might be equated with a hypothetical text (or set of texts) containing information on the topic[85].

There can be no topic without an idea of information on that topic

In fact, topics may be considered cognitive constructs intended to delimit relevant information into reasonable chunks.

Topics are cognitive constructs intended to portion relevant information in reasonable chunks.

[84] Note that we are not concerned with topics in the sense of information structure, which is something completely different (Alfort, 2007). For our purposes, topics pertain to entire texts or passages.
[85] Note that *text* in this sense covers any stretch of language conveying information.

This formulation hints at the next surprising truism, namely that there can be no “reasonable chunks” without there being someone to whom they are reasonable, yet topicality and aboutness are normally considered objective, system-oriented kinds of relevance – a simple matter of matching labels (Cooper, 1973; Janes, 1994; Cosijn & Ingwersen, 2000).

> Topicality, the relation of a document to the topic of a user’s query, (…) really does not depend on the user at all. Determinations such as this could presumably be made by subject experts, intermediaries, or others. (Janes, 1994, p. 161)

Topicality is hence frequently assumed to be relatively easy to emulate.

> Topicality in the sense of matching system artifacts like subject terms is valuable primarily for its practicality: its operational applicability, observability, and measurability. (Schamber, Eisenberg, & Nilan, 1990, p. 759)

The notion of matching topic labels originated with the ancestral library systems where all documents had topic labels and the query was compared to these. Such labels are in fact an instance of *topic-as-thing*, and the kind of relevance at work is consequently indeed the objective or *algorithmic* kind that is meant to be completely context-independent and consequently is highly unrealistic in practice. This does not really have much to do with topicality in the sense of finding information on a given topic, however. In order to distinguish true topicality from this lifeless imitation, Borlund (2003) names the former *intellectual topicality*, the idea being that the user must assess whether a document is truly relevant to a topic instead of just letting the system perform a string comparison. These two types of topicality are frequently confounded in the literature, probably because the nature of topics is somewhat elusive and very abstract.

While the relevance of information on a certain topic to someone in a certain kind of situation must be considered subjective and the ultimate decision left to the person in question, it is easier to accept that information on certain topics related to the main topic might be considered objectively relevant as soon as that central topic has been selected. After all, isn’t information on RIDING always relevant to the topic of HORSES? This is an ontological link between concepts perceived to exist in the real world, and consequently to be objectively true. Objective relevance assignable by subject experts derives from the ontological relationships between topics that can be generalized and are therefore publicly accessible (Kemp, 1974). These relationships are partially objective and indisputable, but even ontological information differs according to people’s culture, knowledge, and even interests, and as we shall see below the topic of horses is certainly not similar across situations of use; some might in fact consider

ASSES and MULES to be much more intimately connected to the topic than RIDING, and while British people might consider DOGS and FOXES to be related to the topic of HORSES, these are certainly not relationships that a Dane would normally think of, because they are contingent on culturally specific behaviours and activities. In other words, the delimitation of topics is fundamentally subjective and the reason that experts and others are able to judge topicality in advance is that some associations are more conventional or even natural than others, but none are truly objective.

Even so, it is just possible that a working system could be constructed on the basis of user profiles while letting the user adjust the weights assigned to ontological relations slightly whenever needed. This is not something that we shall be attempting here, but we shall look much closer at the nature of topics and topicality, because there is no doubt that they are at the centre of the phenomenon of relevance, and in spite of the fact that they have been discussed intensely for decades no real understanding has been achieved. We shall consequently start by looking at the question from a cognitive point of view and see what can be made of it in IR terms.

4.5.1 Topic domains

A query in its simplest form is often believed to represent what is generally known as a *topic*. This is probably to be understood not so much as what the query is about as that which the desired documents are to be about. However, for a document or piece of information to be relevant to the topic-as-thing (the label) is equivalent to being relevant to the query. The reason that we may feel that a document on HORSES is objectively relevant to the query *horses* is accordingly that it contains information on the topic-as-domain *corresponding* to the topic-as-thing that is the query. As with objective relevance, topic-as-thing is not a very useful notion in information retrieval; it is doubtful whether topics can indeed exist without a subject. If the document is on the topic mentioned in the query this is only relevant if there is a user who by submitting the query intended to identify information on that topic. In other words, while relevance has often wrongly been explained away as depending on topicality it seems in fact to be the other way round; topicality is fundamentally based on the concept of relevance and is a by-product of the human cognitive capability of relegating irrelevant information which is not part of the current sphere of interest to peripheral status (see section 4.6).

A topic consists of information which would be likely to be relevant to a person interested in that topic.

This definition of topics is clearly circular, and it will have to be circular because the notion itself is circular: The topic HORSES contains any information that you might find interesting if you are interested in the topic of HORSES; it cannot be reduced to a non-circular concept. The boundaries between two neighbouring topics are not fixed, nor even objectively defined – they are formed by the fluctuating bounds of people's interest spheres. Thus, topics are anything but objective, and highly dependent on the notion of relevance[86].

Relevance to a **topic domain** implies that either the information that is relevant is included in the topic domain or the domain can easily be extended to include it. Thus, information on OATS is typically (peripherally) relevant to the topic of HORSES since the HORSE domain can easily be extended to include OATS. On the other hand, information on STAMPS is usually less relevant, because even though the horse domain *may* potentially be extended to include STAMPS, this would result in an extremely broad domain which would exclude very little, and hence such a construction would not really be helpful in delimiting information that it is worth the effort to acquire from the masses available[87], which is the point of such delimitation of interest spheres, witness Sperber and Wilson's comment that "cognitive efficiency in humans is primarily a matter of being able to select, from the environment on the one hand, and from memory on the other, information which it is worth bringing together for joint – and costly – attentional processing" (2002, p. 14). In other words, the size of topic domains is pragmatically determined. It is therefore hardly surprising that identifying relevant information should be such a difficult task.

Let us return briefly to the example discussed by Cooper (1971).

4.3	**halogen elements**	*

The model proposed here deals with this case in the following way: Users may or may not include water in the topic alluded to by the query, depending on their understanding and current construction of the concept, but they are in any case unlikely to think of halogen elements when reading about water in the U.S. constitution. In other words, in this case users presumably do not include halogen elements in the

[86] Note in this connection that topics differ from *concepts*; while it is indeed possible to make a reasonably clear objective definition of what should be included in the concept HORSE, this does not hold of the topic, because it includes so much apart from the item itself, the reason being that topics delimit *information* and not entities. Some users of search engines may be looking for mentions of certain concepts such as HORSE, whereas others would be interested in the *topic*. It is the latter type of search that we are attempting to deal with here, and it is by far the most difficult to handle.

[87] To be sure, an *ad hoc concept may* be made to include both HORSES and STAMPS without necessarily involving much else (Barsalou, 1983). However, such a concept would be unlikely to be triggered by the term *horse* on its own.

topic reconstructed for the document, and consequently the topics of query and resource do not match. In a would-be omniscient inferential system such differences in construction would be hard to simulate because a computer considers every possible relation unless told not to.

We may conclude, then, that topics depend on a judging subject (i.e. a speaker or user) to define their extent. Consequently, it makes no more sense to posit that a resource is relevant to a topic without meaning this to imply that it is relevant to the user than to claim that it is relevant to a *query* without such reference.

Relevance to topics or queries is meaningless except as a shorter way of saying that the resource is relevant to the speaker or user.

Even in the simplest of cases, such as when the term *horses* in a document is said to be on the topic of HORSES, this is not as objectively true as it might seem. It is of course extremely *likely* that it will be "relevant to the topic", but not by divine law. In the right context, *on the topic of horses* may in fact mainly imply the topic of HORSE RIDING, with the actual animal only occupying an auxiliary role, and WILD HORSES and ASSES being patently beside the point, in spite of their being ontologically subordinate concepts relative to the most general sense of HORSE[88].

A foreseeable objection to this extreme subjectivism is that some delimitations of topics seem so much more natural than others that there must be some kind of objective reality behind them. Some topic domains are more or less culturally defined due to a tendency to delimit topics in similar ways within a homogenous society. Others are ontologically triggered by a natural co-occurrence of some phenomena rather than others. That is, a class consisting of HORSES and SHOE POLISH is not likely to be defined by anybody as a topic. This, then, is where social and cultural relevance occurs. When authors such as Hjørland (2010) claim that it is possible to assign topic labels to resources, this is what they mean: To the extent that topics are culturally defined or ontologically consistent, it is indeed possible to do so. However, there is absolutely no guarantee that a given user will adhere to such conventions; topics may always be subjectively defined by the user.

[88] Relevance between the query *horse* and a document on horses is furthermore only objectively present to the extent that it is objectively true that if a person utters the word *horse* he does in fact refer to a horse. This is not necessarily so, though we would be exceedingly surprised if he were to deny having referred to a horse. Even so, in the end, only the speaker can claim to actually know what he or she was referring to; it is just that it is extremely *likely* that the word *horse* implies reference to a horse. But this is crucially not the same as objective truth; the term may for instance have been used in a metaphorical sense. Be that as it may, in such cases, it is not unreasonable to treat the relevance as for all intents and purposes objectively present.

The trichotomy of personally, culturally and ontologically defined topicality corresponds to what Karl Popper called World 2, World 3 and World 1, respectively (Popper, 1972; 1978). In the end, the personal mental structures define what a topic is going to include in a certain situation, but these may be heavily influenced by cultural and ontological considerations. The psychologically based but physically expressed World 3, Popper argues, may influence us just as much as the physical World 1 itself. Language is inherently part of World 3; through it ideas and constructs formed in one mind are communicated to others. Conventional definitions and conceptions are likewise part of World 3, whereas personal ideas, whether influenced by the physical world or the authority of others, or completely idiosyncratic, represent World 2. Clearly, the delimitation of topics can be influenced by all three worlds.

Note finally that full questions trigger (pseudo-)objective topicality, because we can be reasonably sure about what is a relevant answer to a question; in this case we simply rely on discourse relevance. This is useful in question-answering systems, as well as in inferential systems such as the one envisaged by Cooper (1971).

4.5.2 Topics and genericity

There is generally thought to be a close relationship between topic labels and generic terms describing types, because topics mostly concern generalized types (Behrens, 2005), and indeed queries also frequently – but far from exclusively – contain generic references. Like generic expressions, topics are idealized generalizations over possible situations with certain specified characteristics (Alfort, 2009). They may be characterised by the occurrence of a type of entity, or of a type of situation or event. It was suggested by Alfort (2009) that, contrary to the way they are generally represented in predicate logics[89], generics are not in fact generalizations over the members of a type but rather over perspectives on possible relevant situations including such members. Thus, the generic term *horses* does not refer to all members of the type HORSE, but rather activates an idea of all possible relevant situations in which horses might occur. If it were a genuine generalization over members, it would be expected to cover all horses by definition, but this is well known not to be the case; generics allow exceptions because only relevant members are included (Kadmon & Landman, 1993). The relegation of the generalization from individuals to perspectives explains this seeming irregularity, because perspectives are assigned by speakers and they automatically exclude all that is irrelevant[90].

[89] Cf. a representation of *horses* as $\forall x$ horse(x), which explicitly insists on generalizing over *all* instances of the type.
[90] The notion of *perspectives* in this sense is introduced properly in chapter 5.

Like all situations do, those over which these generalizations are made naturally have contexts, and hence topics, too, have associated contexts. The boundary between the central topic and its context is not always clear, consisting as it presumably does in the distinction between concepts stored mentally with the label and other more loosely associated material. Topics reach out into context in vague and indefinite ways; they even reach into one another, creating links between related topics. Consequently, it is possible to navigate between related topics, and incorporating this in the interface of a LAT is likely to be a good way of evading the so-called *anchoring effect* by forcing users to reconceptualise their information needs (Ingwersen, 1992, p. 129). Navigation may of course not only be situational but also categorical, via a taxonomy of related concepts. Notably, the two methods will not always create the same judgments of relevance. Thus, a SAW is clearly not a TREE (and hence a document on SAWS is irrelevant to the topic of TREES in a categorical sense), but the two phenomena do co-occur ontologically in some important situations and therefore documents on SAWS may in certain specific contexts be relevant; the user might for instance be confronted with a tree which needs removing[91]. Also notable in this connection is the *basic level* effect (Murphy, 2002; Alfort, 2010). If you issue the query term *birch*, your request is too specific for a document concerning SAWS to be relevant in most cases; saws are simply more closely related to the general concept of TREES than to any species of tree in particular.

Subordinate concepts is one area where you might expect absolute objectivity. However, if a document refers to entities of a superordinate type mentioned in a query then this is only relevant if there is a user who intended to find documents mentioning such entities. For instance, a document on PENGUINS might be thought to be relevant to the query *birds*, penguins being undeniably birds, but this is only the case if the user intended his query to refer to penguins as well as to more prototypical birds. *Birds* being a generic term, it inherently involves the exclusion of any irrelevant members of the ontological class.

4.5.3 Domains and conventionality

While it may be difficult to determine what is relevant, there are clearly resources that fall outside of the user's domain of interest. In fact, such domains have much influence on how queries should be interpreted since vocabulary, ontological relations and likely intentions all depend on the domain. Consequently, a LAT should be tuned to one domain as far as possible. In web search this is difficult, because all kinds of domains must be catered for, and people furthermore search across domains

[91] This particular example may seem far-fetched, but in more technical cases users often describe their problematic situations in the hope of receiving advice on what tools to use in order to solve their problems.

(whether they want to or not). In enterprise search, on the other hand, vocabulary, ontologies and query interpretation systems can be tailored to suit a particular domain.

Domain awareness is known to be a good way of disambiguating queries; if the user is a biologist, queries such as *jaguar*, *python*, *apple*, *palm* and *java* are likely to have quite different meanings from what would be the case had they been submitted by a software engineer. Any successful disambiguation of course presupposes semantic indexing in order that the system may know what the terms refer to in the documents that it is supposed to retrieve. Also, information on the user type is needed; if the LAT is located in a site pertaining to a biological society, it is likely that users are more interested in the biological versions of the terms[92]. Otherwise, previously used queries may be consulted for domain analysis (Daoud, Tamine-Lechani, & Boughanem, 2008; Li, Sun, & Datta, 2011); if there are more biologically oriented queries than software-related ones in the user's query log then it is likely that he or she intended to search for the biological terms.

Domains are closely related to topicality in that the right to define a domain may frequently be claimed by people engaged in work related to a certain topic. Thus, a domain called BIOLOGY properly covers information conventionally believed to be relevant to biologists and other people interested in biological issues. For this reason, domains and topics are regularly confounded. This leads to statements that topics are conventionally defined by domain experts, when in fact it would rather seem to be the case that though domains are conventionally defined the topics relevant to their adherents are not. What is relevant to people belonging to a certain domain may change with the development of scientific knowledge within that domain; old issues may become irrelevant and new ones may arise[93]. A document does not cease to be on a certain topic because it falls into disuse within a community; it may well be considered utterly irrelevant by the community of biologists but still clearly be on a biological topic. Despite this, Hjørland advocates a view in which search is first and foremost domain- or culture-based because topicality and knowledge are social constructs.

> People in general act according to how they see the world and what they trust. The first is about theories of reality, the last is about theories of knowledge. When we grow up in a culture, we acquire – both implicitly and explicitly – theories of both kinds. (Hjørland, 2010, p. 230)

[92] On the other hand, it is likely that the terms in question will mostly have the biological interpretation in the documents anyway, so that there is nothing to disambiguate.
[93] This is captured by the criterion of *timeliness* used by some authors (Bruce, 1994).

In other words, people adopt a certain understanding of every concept with which they interact, and often this understanding is coloured by the received opinion within a domain. This is certainly true. However, it is well known that people do not always adhere to such domain perceptions. Thus, the same person may well consider a TOMATO to be a BERRY in scientific terms but still treat it as a VEGETABLE in everyday life. Consequently, domains are adopted whenever they are relevant to the task at hand. They are not given for a certain topic – not even for the same user. If a search is known to be performed by a user belonging to a domain such as BIOLOGY, conventional definitions of what is relevant may be very fruitful; some sources are clearly more relevant than others when it comes to acquiring an up-to-date, authoritative knowledge of a biological subject. However, it should certainly not be confounded with topicality. Users who do not access the topic from the point of view of that particular domain may value entirely different criteria. One example would be a historian researching the perception of animals in Victorian times, and thus specifically interested in sources that modern scientists would consider decidedly out-dated. Another example would be a general user looking for photos of echidnas – even old and unauthoritative sources may well supply useful pictures[94].

The choice of domain-based vs. topic-based search is an important one, and it is one that relies on a thorough knowledge of the users of the tool. The main use of domain-based search is science, and hence some of the most important applications are library tools and sites dedicated to scientific learning. Hjørland (2010) provides a discussion of the theory behind such solutions. With more typical kinds of enterprise search, conventional domain knowledge is likely to play a much less prominent role.

Domains are not just about the topic of a text or a query; they also include a level of technicality (Eggins, 2004). Thus, apart from registering the words of the query, an intelligent LAT should be able to recognize a word as belonging to a certain level of technicality. Paradoxically, this is often not necessary with a simple tool that just matches the string, because if a technical term is submitted, the retrieved documents are likely to be technical too. However, as the tool becomes more sophisticated and includes a thesaurus with synonyms and an ontology with related concepts, there is no longer any guarantee that these belong to the same technical level. Consequently, it becomes necessary to store the level of technicality in the ontology or in the lexicon in order to retrieve documents that match the required level. It is a recurring theme in IR that every new feature that improves the search introduces

[94] See Choi and Rasmussen (2002) for a study of criteria employed in the assessment of images. The most important criteria identified in their study were *topicality*, *image quality* and *clarity*; not recency or scientific accuracy.

new complications and requires the system to take new details into account that were not necessary in a simpler system[95].

Facet	Queries	Topics	Information	Resources
II	Substantial query (collection of query strings)	Substantial topic Topic-as-thing Query string	Substantial information Information-as-thing Resource text	Substantial document (collection of information-bearing entities)
III	Formal query (collection of query terms)	Formal topic Topic-as-expression Query term	Formal information Information-as-expression Resource term	Formal document (collection of resource terms)
IV	Representational query (collection of query domains)	Representational topic Topic-as-domain Query domain	Representational information Information-as-knowledge Resource knowledge	Representational document (collection of resource knowledge)
V	Interpersonal query (containing material with the ability to convey a topic-as-request message)	Interpersonal topic Topic-as-request Query message	Interpersonal information Information-as-process Resource message	Interpersonal document (containing material with the ability to trigger information-as-process)

Fig. 4.7. The parallelism between queries, resources, topics and information; all relate to the facets of linguistic messages. Only the four central facets are included in this concise overview.

4.5.4 Information and topic facets

In the end, then, even topicality is far from simple. In fact, as pointed out by William Cooper (1971, p. 32), "this matter of topic queries is perhaps the stickiest point to be dealt with in connection with relevance", despite a widespread feeling that topical queries in some way represent the simplest and most basic type of search. They are certainly the type that people tend to think of first, but that does not make them less complex. On a surface level, a topic is a linguistic label (a predicate) referring to a class. At the same time, it is also a semantic extension of that reference to a domain that may be of

[95] Technicality is closely related to the level of expertise of the user, though these are in fact independent variables. Attempting a guess as to the user's level of expertise from the choice of words in the query would be risky to say the least.

interest to certain people. Notably, it cannot be defined in this sense without reference to such people; hence, topics are subjective.

There is an interesting parallelism between information and topics on the one hand and resources and queries on the other. All seem to be conceivable in the system of facets introduced in Part I (fig. 4.7). A resource may convey information on all of these levels, i.e. the substantial resource contains words representing **information-as-thing**, the formal resource contains terms representing what might be called **information-as-expression**, the representational resource contains **information-as-knowledge**, and the interpersonal resource conveys a message triggering **information-as-process**. A substantial query contains words representing **topics-as-things**, the formal query contains terms representing **topics-as-expressions**, the representational query refers to **topics-as-domains**, and the interpersonal query conveys a **topic-as-request** for particular information that is useful in a given situation, such as facts. Finally, the intentional resource conveys communicative intention and the effective resource affects the recipient psychologically, two features of information which must be considered central whether one includes them in the definition of information or not. The fact that in this view they represent facets rather than characteristics of information in general explains why they can be dismissed from the definition of information here while remaining absolutely central to the concept; they each correspond to one of the facets which make up information, and so they are in principle always present to some extent yet at the same time they only describe that one facet and not the entire concept of information. After all, the recipient of the intended resource may be ever so vaguely defined, and the psychological effect ever so slight without robbing a communicated content of its status as information. At the same time, speaking of information without thinking of these two facets as being at least potentially present is unthinkable.

This overview concludes our discussion of topics by clarifying the relationship between topics as labels and topics as domains and relating them to the corresponding facets of information. The faceted character of both topics and information will be essential to our discussions in Part III where much more is said about the role of the various facets.

4.6 Irrelevant information

Although the phenomenon of relevance is now known to include much apart from topicality, the connection remains in place, and is in fact still much too prevalent in the literature as a whole. Nonetheless, topics and matters of interest are certainly fundamental notions that must be dealt with if we are to understand relevance. The former are likely to be central to our cognitive organization of the world as we perceive it, and the latter derive from our essential basic capability of avoiding the irrelevant. This ability allows us to focus on the set of stimuli that really matter rather than drown in the sea of information that our senses are capable of collecting. Indeed, "more and more evidence points to the fact that learning is, in fact, forgetting" (Boeckx, 2010, p. 49), i.e. learning to ignore the irrelevant.

It is an important task of IR systems to fight what has in the field of lexicography come to be known as *information death*, i.e. when users are presented with so much information that they suffer from informational suffocation and end up giving up without even obtaining the important information they came for (Bergenholtz & Gouws, 2010; Olivera, 2010). Thus, ridding the user of the irrelevant is of the utmost importance.

According to Wilson and Sperber (2004), human cognition is likely to have evolved to always maximize relevance by weighing potential gains from attending to some stimulus (and, one might add, losses from ignoring it – this must surely have been of great importance to early hominids) against the processing (or other) efforts likely to be required to do so. It is thus a tendency which is deeply ingrained in our very genes. The suggestion is put forth here that, in IR as in life in general, the notion of *irrelevance* is far more significant than is usually acknowledged; most kinds of information retrieval are basically not concerned with retrieving some stored knowledge of a certain predefined kind but rather about suppressing the large amounts of irrelevant information available and navigating through the vastness of irrelevance until something useful presents itself.

4.6.1 Interest and relevance

In the end, the only thing that counts in search is what resources users want to see and which ones they would rather not bother with, and all other technical considerations of relevance are merely proxies for this[96]. Objective relevance is the simplest way of avoiding too much useless information, but situational relevance is a huge step towards capturing actual **interest**. Interest is closely related to topicality, i.e. what the information concerns, but in addition it relates to the relative importance of various interests to a human perceiver in a given situation. Saracevic (1996) brought the work of the philosopher Alfred Schütz to the attention of the IR community; Schütz distinguished between three basic types of relevance which interact dynamically in all contacts between humans and their environment, or *life-world*, as he called it (Schütz, 1970). These three interest spheres, for which he used the terms **topical**, **interpretational** and **motivational relevance**, pervade the human mind and consequently also guide relevance evaluation in search. In connection with search, Schütz' *topical relevance* corresponds

[96] This is captured in the Regulated Flux model suggested below by making desirability the final criterion of relevance assessment.

to that which is relevant to the task, while *interpretational relevance* relates to the background or long-term interests of the user and *motivational relevance* to the short-term interests represented by the query. Thus, in his terms information is topically relevant if it is helpful, interpretationally relevant if it is merely interesting and motivationally relevant if it is on the topic of the query. Schütz explains that every individual has at any given time a *kernel* of knowledge that is taken for granted and on which further understanding is based. This is surrounded by a *horizon* of less straightforward matter which is potentially problematic or of interest but not necessarily relevant at the moment. It is against this horizon that relevant stimuli – *themes* – are identified with reference to the kernel.

> One experiences that which is taken for granted as a kernel of determinate and straightforward content to which is cogiven a horizon which is indeterminate and consequently not given with the same straightforwardness. This horizon, however, is experienced at the same time as fundamentally determinable, as capable of explanation. (Schütz & Luckmann, 1973, p. 9)

The actual names assigned by him to the three basic relevance contexts are somewhat unfortunate in an IR context, since topicality is generally associated with queries which represent motivational relevance[97] and motivation with what is effectively interpretational relevance. Consequently, we shall take the liberty to rename them for the purposes of IR research in a more appropriate and, it is thought, straightforward way as **focal**, **local** and **general relevance**.

- **Focal relevance**: Something is selected by the individual as being of current interest. (Schütz' *motivational relevance*)

- **Local relevance**: Something is perceived as being problematic and is distinguished from the horizon as a theme on which to concentrate. (Schütz' *topical/thematic relevance*)

- **General relevance**: Something is judged as being relevant on the background of the horizon, the background knowledge possessed by the individual and the past experiences. (Schütz' *interpretational relevance*)

We shall want to make a slight amendment to the definitions as formulated by Schütz in that local relevance should be able to concern unidentified problems as well as known ones; a stimulus that

[97] In fact, the term *topical relevance* was later changed to *thematic relevance* (Saracevic, 1996).

makes the person aware of a problem may be just as important as one which solves a known problem, yet it seems from the definition above that Schütz was mainly thinking of known problems.

An example of the application of these spheres of relevant context to a search situation will be illustrative: If a user searches for railway tickets to Lisbon, resources on RAILWAY TICKETS are clearly relevant to him in the sense of focal relevance – he chooses to be interested in that subject for the time being. However, at the same time the lower price of plane tickets may make resources on those highly relevant in the sense of local relevance – plane tickets might well satisfy his task needs even better. Finally, if the user happens to be a fanatic boat enthusiast, sailing to Lisbon might in fact turn out to be the ultimate way of getting there, and hence resources on BOAT TICKETS would be highly relevant in the sense of general relevance, in spite of the fact that the journey would take much longer. In this light, there is not one kind of relevance which is "better" or more sophisticated than the others; they simply complement each other as covering different aspects of our cognitive interaction with the world that surrounds us. Hence, relevance to a task as well as to long- and short-term interests should all be featured in an IR system. However, it is important that they act in an interdependent way, since long-term interests should usually only be considered if they are in themselves made relevant by the short-term interests. In other words, the fact that our user from before is besotted with boats is quite beside the point if at the moment he is looking for information on beetle reproduction.

4.6.2 The hierarchy of relevant contexts

It is vital for user profiles to avoid falling into the trap of storing incidental information in the long term and colouring all results with them. Exactly how long-term interests are made relevant by short-term ones may well turn out to be one of the great challenges of future IR research. What makes *boat tickets* a reasonable suggestion when *flight tickets* have been requested is that they are both able to satisfy the same task-induced need. Thus, general relevance (long-term interests) not only depends on focal relevance (short-term interests) but also on local relevance (task-induced needs). Only when they all play in concert should the long-term interests be included in the algorithm.

General relevance should not be exploited when contradicted by local or focal relevance.

Similarly, local relevance should not (and normally would not) be exploited when it contradicts focal relevance, i.e. when it goes against the user's direct request.

Local relevance should not be exploited when contradicted by focal relevance.

This emerging hierarchy, which we shall henceforth call the **hierarchy of relevant contexts** (**HRC**), accounts for the order in which the three Schützian contextual spheres are presented above and has wide-ranging consequences for IR system design and research.

Adhering to the hierarchy is not as self-evident as it might seem. Phinitkar and Sophatsathit (2010), for instance, explicitly warn against colouring all results with long-term user-profile information.

> A prevalent shortcoming of search process is that many personal search approaches often return search results by focusing on the user's interests and preferences rather than on user's queries. (Phinitkar & Sophatsathit, 2010, p. 210)

Yet they fall into the pit themselves because if there is no match between a word in the query and an element in the user profile then their system will expand the query by adding the highest-ranking word in the profile. Thus, a word which is entirely unrelated to the query may automatically be added, which must surely confuse the system rather than produce a helpful result. Another such problematic system is suggested by Kajaba and Navrat (2009). This system uses co-occurrence of terms to establish that Miles Davis and John Coltrane are possible substitutes for Herbie Hancock in that they are similar and often occur together. However, surely if the user was searching for *Herbie Hancock*, resources on the other musicians are not equally good results? They should at least be demoted to a much lower place in the results list. This is not to say that the methods advocated in these works are not useful, but that they must always be guided by the HRC.

Even though the HRC should not be violated, what we *can* do in order to help users and avoid the constraints of only presenting information that they have specifically asked for is to convert task needs or long-term interests into matters of current interest. For instance, a task can be objectified by letting users choose the need-generating situation they are in from a list of suggested situation types (or in another more intuitive visual form). In this manner, the task effectively becomes the user's short-term interest, and the matter can be resolved as a case of focal relevance. Long-term interests can likewise be promoted to current interest status through systems that allow users to explore interesting content without necessary reference to a query. It is thus important to either respect the hierarchy of basic relevancies or objectify needs and interests in such a way.

The reason for the hierarchical organization of Schütz' relevancies as applied to IR is clear if seen on the background of Relevance Theory. Because search depends on a linguistic input (a query), this acts as an *ostensive stimulus* and would, had the recipient of the request been human, have generated an expectation of optimal relevance. Since it would seem that on this point we communicate with LATs largely in the same way as we would to our equals, we expect the recipient of the request to interpret a query as being such an ostensive stimulus, and hence to react on it rather than on background information or task-generated needs. Otherwise we are sorely disappointed and consider our communication to the system to have been unsuccessful. This is one of those features which are not adjusted to the man-machine communicative situation.

4.6.3 Irrelevance in cognition

The ability to react to changes in the environment (*stimuli*) and reason about the appropriate way to interact with it has always been one of the great mysteries of life. The power of cognition was long thought to be possessed by man alone; today, it is hardly necessary to add that animals constitute cognitive systems as well. From there it is natural to proceed to all living organisms; even artificial systems are considered cognitive according to some definitions. As shown by Gershenson (2004), all of these definitions can be included in one very general understanding of the concept; they all have something in common, and there is no reason to stay with any one narrow definition, in the same way that this is true of *information* and *relevance*.

Following the Chilean biologist and philosopher Humberto Maturana who first described the concept of *autopoiesis* and coined the term, we shall adopt one such relatively broad definition. According to Maturana (1980), a cognitive system does not require a nervous system but rather is defined in a broader way.

A cognitive system is a system whose organization defines a domain of interactions in which it can act with relevance to the maintenance of itself, and the process of cognition is the actual (inductive) acting or behaving in this domain.

Intuitively, one may be hard put to accept that entities without brains may possess knowledge. However, Maturana assures us that this is merely a question of the degree of complexity allowed by the system in question. Some of the knowledge that we possess is hardcoded in our genes (*instincts*), and this is true of animals and plants as well, and, it can be argued, even by dead objects; it is just that in this case there are no genes, but there is still information in the molecules, and objects have the ability to react to stimuli in certain specific ways.

Does a tree know how to blossom?

Maturana's definition implies that it does; not in the sense of being able to reflect upon this knowledge like a rational being, but that is just a question of the larger complexity that a brain allows[98]. The interactions of a cognitive system with its environment may in fact be purely mechanical or chemical if they are coded in the physical structure of the system, as when a carnivorous plant seizes its prey; or they may be coded in neurons or computers. Coding simply expands the domain of interactions of the organism by allowing interaction with abstract *pure relations*; it is not a required feature of a cognitive system, according to the definition given[99]. This view integrates with the understanding of information presented above in that information is simply more strongly informative when interpreted by a more complex cognitive system such as a human being, but it is still information if merely interpreted by a computer, because that too represents a cognitive system, if a much simpler one. The tree is also able to process information in this sense, being a relatively simple cognitive system (as compared to beings with brains).

Relevance is arguably the most basic of all cognitive processes, and indeed is the pivotal feature of the definition of a cognitive system as given above; a cognitive system is one that is able to ignore irrelevant differences between instances of interaction. In order to understand relevance, therefore, we must understand the basics of human cognition and especially the ability to ignore the irrelevant.

A crucial feature of cognitive systems is their circularity; they can only react to identical states arising on separate occasions and triggering the same reaction, such as successive flies landing on a carnivorous plant (Maturana, 1980). Of course, every interaction is different from all others in some way (for a start, it will always have occurred at a different point in time), so in order for the system to remain in place, all minor differences must necessarily be ignored. For instance, certain carnivorous plants treat all objects of the approximate size and weight of a fly as belonging to the category PREY. If every fly were treated differently by the cognitive system, no predictions could be made and interactions with the environment would be entirely erratic. On the other hand, if the stimuli are sufficiently alike, they are treated as for all intents and purposes identical, and the same response (or range of possible responses) is activated. This necessitates a function that allows the cognitive system to ignore the irrelevant

[98] Recently, *plant behaviour* has come to be accepted as an important biological area of study. Scientists have discovered that plants are in fact able to react to changes in their environment to a much greater extent than hitherto believed (Information and literature can be found at The Society of Plant Signaling and Behavior on www.plantbehavior.org).

[99] Note that even if a different definition were adopted in which coding was required, computers would still pattern with animals rather than with the brainless plants. Maturana's definition should hence not be rejected out of hand simply because one would like to consider cognitive abilities to be a special privilege of animals. It should be kept in mind that cognitive abilities are not equal to the possession of a free will (Searle, 2001).

differences, and that function seems to have been perfected through evolution in human beings as well as in plants (Wilson & Sperber, 2004).

The world abounds with information, most of which is quite irrelevant to us, and it is likely that most of our concepts are accordingly negatively defined. Since no two instances of any class are exactly the same, it is simply not possible to establish a class without in some way defining the ways in which members are allowed to deviate, thereby effectively delimiting the irrelevant variations on the theme (as with the fly). We know what a HOUSE is certainly *not* like and usually recognize a NON-HOUSE quite clearly when we see it. On the other hand, defining a HOUSE may be rather more difficult. Certainly, more than one definition would be needed in order to cover all instances in which the word is used. In many cases we are required to disregard quite substantial differences in order for our classifications and the inferences based on them to work. In fact, it is unlikely that there are ever clear rules as to what is a correctly identified member of a (linguistic) class such as HOUSE. This is not to say that there are never clear criteria for belonging to a certain type in the world (it is hard to deny that there is such a thing as a workable definition of a FISH as a family of species represented by individuals in nature for instance[100]), but classes as perceived by humans are simply not the same as types in the world and may be defined in completely different ways. Even in cases where we do have clear definitions these tend to belong in certain domains such as scientific classification or legal labelling. Outside of these domains, classes can be ascribed rather more freely and dynamically. For instance, a biologist may insist that bananas and tomatoes are berries when teaching her students and nonetheless call them fruits and vegetables as soon as she goes to lunch. In addition, there is always the possibility of using labels metaphorically. Consequently, even though a concept such as FISH does have a strict scientific definition, there is no limit to what can in practice be called a *fish*. One can always come up with yet another thing which might in some context be called a *fish* in a metaphorical sense, given the right kind of situation. One might, for instance, speak of a boot hanging from a fish-hook as being *one strange fish*. What this means is that for something to be classified as a member of a certain class (such as FISH), the observer need not merely identify necessary and sufficient features judged to be relevant, but must above all avoid the irrelevant candidates, which might in some contexts be classed as fishes, but whose inclusion in the class would not in those particular circumstances be of advantage (or in the words of Wilson and Sperber would not create a *positive cognitive effect*). Thus, if you are looking for fishes, attending (passively) to the features of all currently-not-fishes is in fact at least as important as pinpointing those individuals that do belong to the class. First of all, it would make no sense to check

[100] That is not to say, of course, that everybody would agree on exactly where to place the borders between FISH and NON-FISH; even in this simple case there is ample room for controversy.

whether everything you see is a fish; you need to filter out all the unimportant stimuli that are clearly not fishes before being able to concentrate on finding some. Even then, not all fishes which are clearly members of the FISH category in a scientific sense would qualify in your search for marine food; dead and sick fishes, poisonous fishes, fishes made out of plastic, too young fishes and too large fishes, etc., might not belong to that concept at the moment, because the concept is locally defined relative to your needs in a particular situation[101].

The idea that the ability to filter out irrelevant information is crucial to survival is also behind Martin Heidegger's concept of *Geworfenheit* (or *thrownness* in English); we regularly find ourselves performing actions which we are not aware of actually carrying out, or at least do not need to think about how to do (Heidegger, 1996). Some activities simply come naturally to us, having performed them sufficiently often before. Life is a constant process of filtering out irrelevant stimuli that we can deal with passively or not at all, and search does not differ in essentials from other aspects of life.

The central importance of irrelevance is due to the fact that no inference can be based on concrete instances; inferences always concern generalized classes of recurring interaction with the environment. The first time a carnivorous plant meets a hummingbird it must either classify it as belonging to a previously known category, such as PREY (along with the flies), or fail to interact with it altogether; it cannot simply establish a new category with which to interact in a different way. This is not just because evolving new behaviours in a plant takes time, but simply because if there is no recurrence, there can be no prediction. In time, after many encounters, the plant may learn that hummingbirds constitute a different category from flies and require different reactions. It is this process that is so much faster and more effective if you have a brain to process the information; which leads us back to the claim that the development of the brain has to a wide extent been guided by the need to adapt quickly to changes in the environment, or in other words to assess the relevance of items that do not immediately register as a recurrence of a known event or stimulus. It is the same function that is activated when people assess the relevance of a resource – some stimuli are clearly (or almost certainly) irrelevant while others need to be subjected to further scrutiny in order to determine whether they are relevant or just somewhat similar to a relevant resource. In both cases, the central task is drawing the limit between relevant and irrelevant variations on a theme.

4.6.4 Irrelevant information and information death

The question is how to apply these findings to the design of IR systems and other LATs. A distinction should be made in this connection between identifying the relevant so as to guide users of IR systems

[101] See also Kadmon and Landman (1993) and Winograd and Flores (1986), who provide similar examples.

towards their goals, and removing the irrelevant in order to relieve them of information overload. Both of these processes are necessary, but their relative importance varies with the type of search. From the user's point of view, the natural thing to do given the basic organization of the human cognitive abilities is to start by filtering away irrelevant noise, then concentrating on the positive features (relevance) and finally proceed to filter away the remaining irrelevant instances among those tested. In IR, this amounts to performing a rough first partitioning (Blair, 2002a) followed by a search guided by relevance and finally letting the user peruse results and perform some post-filtering in order to remove various types of irrelevant results. This may be suggested as the rough template of search that one ought to approximate. Currently, most effort is invested in the middle section, whereas the partitioning and the post-filtering parts are somewhat neglected areas of enquiry.

It is true that people usually search for relevant material and are only occasionally aware of the need to avoid the irrelevant, but the fact that this is how people go about their searches in no way guarantees that it is optimal, or that search engines should operate with exclusive reference to relevance. Perhaps the fact that what we have called *meta-awareness* is not an inherent feature of search but rather an extra feature so to speak explains why most researchers focus solely on relevance and ignore the importance of irrelevance (see section 10.2).

To be sure, sometimes irrelevance *is* taken into consideration in statements on the purpose of an IR system, but relevance seems always to be mentioned first and to be considered the primary force. Thus, according to van Rijsbergen (1979), "the purpose of an automatic retrieval strategy is to retrieve all the relevant documents at the same time retrieving as few of the non-relevant as possible". In machine learning, on the other hand, it is now recognized that irrelevance is the basic phenomenon, and that relevance is derived from it by negation (Pearl, 1988; Bell, 2000). The idea behind this is that if an inference can be made without reference to a certain piece of evidence, that evidence is irrelevant to the conclusion. In IR terms this means that if the user can (relatively easily) acquire the needed information from another resource (or other resources), then the unnecessary resource is in principle irrelevant. The unknown variable then becomes what constitutes needed information (and also in what form the user prefers to receive the information).

It was pointed out above that the ultimate task of a search engine is to provide the user with useful information, but an important step towards this is reducing the number of documents that need examining to a manageable set.

The primary task of a search engine is to identify irrelevant information so that it can be hidden from view in order to facilitate navigation among the remaining potentially relevant resources[102].

Note that a very narrow domain will increase the average relevance of all documents to any topic within that semantic field. Thus, identifying completely irrelevant information becomes harder – given that there is less evidence on the basis of which to discriminate it – but perhaps also less critical, in enterprise search than in a broad web search environment.

The claim put forth by Schamber, Eisenberg and Nilan (1990, p. 774) that "the field has come a long way from the earliest experiments in evaluating information systems when the objective was primarily to suppress nonrelevance" may seem to contradict the suggestion made here. However, the change from avoiding irrelevant documents to focusing on identifying the relevant ones is in fact not what has advanced the field; it is only a by-product of a deeper transformation. In the early studies, it was thought that a document was relevant to a topic if and only if the topic was mentioned in it. Hence, the cases when this failed had to be avoided as "false drops"; clearly, documents retrieved but rejected by the user were not really on the topic in question since they were not considered relevant. This may or may not have been true in particular cases; however, what brought about the decisive change was the recognition in more recent works that topicality in this objective sense is not everything. When relevance started to be considered a phenomenon initiated and constructed by the user, irrelevant documents returned by the system could no longer be said to necessarily be errors. Rather, the system's reconstruction of what the user would find relevant might be wrong. Hence, a strong focus was laid on identifying the relevant, and the irrelevant was forgotten. However, the fact that relevance is subjective does not mean that a resource is *never* irrelevant (or at least of too little relevance to be of any interest to the user). Consequently, it is time that irrelevance was reintroduced as a central concept in IR[103]. It is in fact only now that we can cautiously begin to make such suggestions in earnest, because dealing with the irrelevant information is likely to require far more processing power on the part of the system, as the amount of irrelevant resources is in general far greater than that of relevant ones. However, as pointed out by Blair (2002a), there may be ways of reducing the efforts required if the focus is shifted and a rough initial partitioning is performed before the fine-tuning search is set in, as suggested above.

[102] This statement should be qualified if other types of search than informational queries are considered.
[103] To be sure, there *are* cases of search where relevance is the driving force because there is only one "correct" result, such as when people navigate to a specific site by entering its name or a part of its URL into the search box, but these are special cases.

5 Naming the topic

> ‘So…er, where is the professor, in fact?’
>
> ‘Oh, in the jar, for a certain value of “in”’, said Professor Pelc. ‘It’s very hard to explain to the layman. He’s only dead for –’
>
> ‘– a given value of dead?’ said Moist.
>
> ‘Exactly! And he can come back at a week’s notice.’
>
> Terry Pratchett: *Going Postal*

We now turn to a discussion of how language is used in search queries. Although there are clearly differences between using language in search and in other contexts, the same basic psychological processes apply, and similar (if not always identical) expressions result; users supply single words, syntactic constructions, and occasionally entire sentences as the input to the search algorithms. In order to be able to determine what intentional differences a particular variation in query expression reflects it is necessary to understand how query terms work in the first place and what their relationship is to the user and to the phenomena that they refer to.

Every (semantically non-empty) word has a meaning, i.e. it refers to a class of some kind, and it also refers to some entity, state or event described. In addition, every complete construction refers to some potential situation (or an aspect of one) in a hypothetical mental space, whether actually occurring or completely imaginary. Section 5.1 deals with the fundamentals of reference in this wide sense, while section 5.2 introduces the notion of *bisubjectivity* and discusses the cybernetic character of language as used in search.

Another inescapable feature of language is perspectivization. Reference inevitably involves viewing a scene from a certain point of view, and this can be the spatiotemporal location of the observer (the default choice) or it can be a different location chosen for the purposes of the portrayal of a situation. This is true of search, too; it is possible for users to search for information needed by someone else, or needed by themselves at another point in time. Apart from this, perspectivization also involves the choice of a scale on which things are portrayed. We have a natural human perspective on things, but there are still wide margins where it is possible to describe something as being of a specific or more general type, or as constituting an individual, a collection of individuals, or an unbounded mass. This linguistic choice is founded on cognitive defaults and subjective choices alike and is highly significant when attempting to reconstruct user intentions. The subject is treated in section 5.3.

The most basic type of reference in search is generally considered to be a topic label. Topic labels are generally associated with generic expressions, and one might think that these were examples of perspectiveless expressions. Not so. I have argued elsewhere that generics represent a generalized idealization over possible perspectives, and as such constitute a very special type of perspective (Alfort, 2009). This is briefly touched upon in section 5.4. Space considerations prevent us from going into great detail in this thesis, but a basic understanding of the processes underlying the use of language in queries is requisite. The following is thus a highly compressed version of the theory of reference adopted here. The reader is refered to the sources for detailed information.

5.1 Reference – The essence of languaging

Just like relevance, reference is a controversial phenomenon, and an extremely interesting one at that; it is the process that allows us to portray situations from a real or imaginary world. Language consisting of mere words, it is close to miraculous that we can use them to conjure up worlds in the minds of others. We choose words, or more technically **predicates,** to describe entities, states and actions in the situation we want to portray, partly from their ability to describe them properly and partly from subjective choices – the same entity can always potentially be described in more than one way. The meaning of a given predicate is defined (often silently and unknowingly) by the language community as a whole, so we can be comparatively confident that the addressee has roughly the same idea of what type of entity it refers to as the one that we intended. On the other hand, we cannot be sure that the addressee would agree that the entity referred to belongs to this class, or is aware that this is so.

We furthermore apply a certain **perspective** on things in order to portray a situation from a certain angle rather than others, and this is again a subjective choice, though some perspectives are more natural than others. The perspective chosen has repercussions for the choice of predicate and the level of generality or specificity of description, as well as any indexical, deictic, local or temporal expressions used. In verbs, it affects tense, aspect and modality. Having received these types of information through the linguistic expression, a hearer will then hopefully be able to reconstruct the situation from the appropriate point of view, and even infer a range of other possible perspectives on it. This reconstruction happens in the mind of the hearer, and indeed the world portrayed by the speaker was likewise a mental model constructed in the mind of the portrayer. What the speaker refers to is thus not a situation in the real world, but a mental construct (Fauconnier, 1994; 1997; Dik, 1997; Alfort, 2009).

> The notion of 'reference' is well established in the current literature (although rarely defined carefully!) as having as target the 'underlying entity' (who or whatever it is), and as signaling to the hearer who or what it is that I am speaking about. Our definition does not change this, but adds the requirement that the individual denoted be referred to under a certain aspect or aspects, and it is these aspects that determine the meaning of the reference[104]. This means essentially that our idea of reference has as target the <u>mental representation</u> of the referent, as well as the underlying entity. And if the mental representation of the referent is different, the reference is different. (Riley, 2007, p. 849; emphasis original)

This is true whether the subjective mental world corresponds neatly to the real world as perceived by the speaker or is entirely fictive, or simply slightly modified for various reasons known or unknown to the speaker.

> [T]heories of reference based on the use of language cannot bypass mental spaces; they will have to forsake the idea of a direct link between linguistic structures and referents and take into account the important intermediate process of space construction. (Fauconnier, 1994, p. 158)

Mental spaces is Gilles Fauconnier's term for the subjective mental world constructed by the speaker before portraying it linguistically.

Many linguistic discussions of reference – whether based on Saussurian or Peircean views or of a younger date – somehow manage to include real-world referents in the equation so that signs are said to represent an object in a real world for the hearer to interpret and reconstruct, and this is a dangerous mistake. Here is a traditional example.

> A relation is thus seen to hold between the extra-linguistic entities on the one hand and the linguistic entities on the other: the relation *is referred to by means of* or, as we shall henceforth say, *is denoted by*. [In the sentence "Anderson met a man with a dog"] Anderson is denoted by "Anderson", a man by "a man", and a dog by "a dog". (Sørensen, 1958)

[104] Riley's notion of *aspect* is related to, but not identical to, my notion of *perspective*.

This is probably not a view that many serious reference researchers would ascribe to today, but one still gets the impression that it dominates the general understanding of the phenomenon among people who are not particularly engaged in work on the issue. In fact, strong arguments can be made that there is not even anything out there to refer to until the reference is made and the emergent entity is constructed as a cognitive entity (Winograd & Flores, 1986). After all, the individual things that we perceive are not really there (in the form that we perceive them) but are the result of our observation and cognitive treatment of the input (Lakoff, 1987)[105].

5.1.1 Application and accommodation – The two halves of reference

I have suggested elsewhere that the subjective world constructed in the speaker's mind consists of what might be called **applicants** – entities which can potentially be described by the application of predicates, such as individuals, properties, actions and situations[106] (Alfort, 2010). All applicants have a set of spatial, temporal and classificatory relationships to each other[107]. They differ from all others on one or more of these parameters, and to varying degrees. Furthermore, there are no absolute coordinates in any of the dimensions in the subjective world. This means that situations can be viewed from different angles, but also that objects are not classified in any decisive way; the speaker provides the missing values to these parameters. There is consequently a certain amount of freedom in the manner of description of a given situation, such as the (exact) choice of predicates and the perspective, but once these choices have been made, the expression also has to adjust itself partly to the portrayed world in order to be an

Fig. 5.1. At a prelinguistic stage, a situation is likely to be represented in our minds as a model capable of being viewed from different perspectives, and with no actual labels assigned to the entities involved.

[105] This does not mean, of course, that there is no world out there, but only that it is not organized the way we perceive it; *constructivism* does not imply *solipsism* (Heylighen & Joslyn, 2001). It is just that the world that we refer to is necessarily always the one constructed in our minds when the physical world is perceived and interpreted from a human point of view.

[106] My *applicants* are closely related to Cognitive Grammar's *profiles* (Langacker, 2000). The difference is that the term *applicant* is meant to suggest potentiality – it covers describable entities in the world, to which predicates *may* be applied – whereas *profile* suggests actual profiling of certain facets of the world.

[107] The latter means that they are of different types, but these types are necessarily related in various ways and to various extents. Most basically, every type is placed at a conceptual distance to every other type in a sort of conceptual plane which we shall later speak of as the *classificatory dimension*, cf. section 5.3.1.

acceptable portrayal of that particular world. If the choices made by the speaker are too fanciful, the expression will not be able to match even the subjective world and will be uninterpretable. Whatever linguistic expression we use in our portrayal of a constructed mental world, the lexis and grammar we employ must agree with its structure and features (but not necessarily with those of the real world in front of us). Number and gender agreement must be correct, for a start; otherwise it cannot be considered a correctly formed expression. Thus, if we imagine a situation with two seahorses, we should use the plural when describing it, not the singular. This does not hold, of course, between the expression and the real world, because we may well be mistaken about the number of seahorses in the portrayed situation in the real world, or we may have reasons to mislead our hearer about it, and neither of these conditions would make the expression incorrect in a linguistic sense. Thus, as soon as a mental world has been constructed for the purpose of producing an utterance about it, the speaker is bound to conform to it. Otherwise the hearer would not be able to reconstruct the speaker's mental picture, and we have to trust that the person speaking to us is conforming to his or her subjectively modelled world.

There are two stages at which the speaker has a choice to make; firstly, the mental world that the speaker wishes to portray is modelled subconsciously by the speaker in a subjective way, though influenced by the external world in general and sometimes in particular; and secondly, the expression is constructed more or less consciously by the speaker in accordance with this world. Users are free to express themselves in any way they like, as long as the expression conforms to the model. In order to account for these circumstances, it was suggested in Alfort (2010) that reference involves two layers: The **application layer**, where speakers make the necessary subjective choices to produce a description of the world that they wish to portray, and the **accommodation layer**, where the expression accommodates to the world in order to produce a consistent portrayal of it. The application layer takes care of the mappings between speaker and expression, while the accommodation layer manages those between the expression and the world referred to. It follows from this division of labour that there are two kinds of subjectivity: The **direct subjectivity** consisting of the choices made by the speaker on the application layer, and the **indirect subjectivity** resulting from the subjective construction of a world in the first place.

Whenever a speaker uses a predicate, it applies to all referents which are classified in the way denoted by that predicate. The fact that the predicate as used in the portrayal of a given situation **refers** to one or a number of referents in that context only is determined on the accommodation layer in which the predicate adjusts itself to the world. If there is only one relevant applicant in the subjective world, the predicate refers to that alone, and the speaker must use a singular form in order to produce a correct portrayal of the world.

It is frequently emphasized in the literature on reference that a predicate does not in fact refer, but rather the speaker refers through the predicate (Riley, 2007). I myself have until recently taken this to be a truism (Alfort, 2009). However, it now seems to me that the speaker does not refer directly to applicants after all. Rather, what speakers do is construct the worlds that contain the relevant referents, followed by the application of predicates. Reference then occurs as sheer accommodation to the portrayed worlds. This involves some restrictions on the actual expression of the utterance when it comes to such features as gender and countability in nouns and transitivity and aspect in verbs; the expression must show agreement with the constructed mental world as seen through the adopted perspective.

5.1.2 Predicate and term reference

When used in an utterance, every noun and verb refers to some entity, state or event in the portrayed world. Here, I use the term *refer* in a broader way than is generally the case, to include verbs and what is traditionally known as non-referential nouns, because they too are given extensional descriptions through the words that we use. I have argued elsewhere that this is a reasonable – indeed, necessary – extension of the term, as also suggested by Conrad (1982) and Keizer (2005)[108]. Not only nouns pinpoint their referents in time and space and give them extension in the world; so do verbs. They both denote phenomena which have spatio-temporal extension, the difference being one of degrees between those phenomena which are prototypically relatively stable through time and have physical extent and those which are prototypically ephemeral and non-physical. The term *referentiality* can then be used – as indeed it is at present – to explain the difference between referential and non-referential nouns.

Application and accommodation together make up the phenomenon that I have elsewhere called **predicate reference** (Alfort, 2009). This is not the kind of reference that is normally understood by the term in modern linguistics – that would be **term reference**. The latter phenomenon covers the process in which actual entities are referred to and constructed as discourse entities. Ferdinand de Saussure's lectures are a classic in reference research, but unlike many other reference researchers after him, and often heavily inspired by him, it is clear from de Saussure's notes that he was primarily

[108] It also follows from the formal definition provided by Riley (2007), at least for all nominals, and perhaps also for verbs if he had not explicitly restricted his paper to the study of nominals, cf. "The Principle is somewhat non-standard: what it amounts to is that, in a discourse, all nominal expressions are 'denoting' or 'referring' expressions (although I do not use the latter terminology, and although the denoted entities are not necessarily individuals, but may be more general entities, sets or other logical constructs)" (Riley, 2007, p. 838; emphasis original).

interested in *predicate reference*; to him, a *signifier* referred to its *signified*, which was the type of entity rather than any actual individuals (de Saussure, 1983). As the type does not originate in the speaker's mind but ultimately in the external social reality, the signifier has often wrongly been taken to refer to something external, including objects, as also noted by Slobin (2005). In my view, predicates do *not* refer to external types, but to features of the mental model, using predicates agreed upon collectively on the basis of types that exist in the social reality. The construction of the mental world with its types does not participate in reference in this view, but rather precedes it and makes it possible.

Term reference occurs not with the speaker as agent, but in the hearer's mind, guided by the speaker's utterance. Whenever an individual entity is referred to in the speaker's speech, either a new discourse entity is set up by the hearer (and the speaker), or an identifying referential link is made to an existing discourse entity or accessible contextual entity. Thus, while the speaker refers to types using predicates, the actual term reference is carried out by the hearer under the speaker's guidance. Even overt "reference" to external objects present in the speech situation is made to discourse entities representing them in the "collective mind" or *consensual domain* of the speech act participants (Maturana, 1978; Winograd & Flores, 1986). The fact that there is an explicit link between a discourse entity and the object in the external world is not a feature of the referential act, but of *deixis*, which has a far stronger relationship to perception and to our embodiment as human beings in the world. Deixis is of minor importance in search and hence shall not be discussed further at present.

5.1.3 Applying a perspective – Reference as an observation

The description of something invariably involves someone observing it from somewhere. In order to be able to imagine the existence of an entity in the first place, you need to be a rationally thinking observer. In addition, if you want to be able to register the location of it you will need to be located somewhere in the same world yourself. Its location can now be established relative to your own. Finally, if you want to be able to perceive its properties (what it looks like or smells like, how it acts and so on), you need to perceive it and process the information. Any perspective is thus grounded in what I have elsewhere termed an **observer** and a **host** (fig. 5.2) (Alfort, 2009). The *observer* is a rational being; most frequently this is the speaker, but the observing role may alternatively be transferred to another rational subject who observes and evaluates the makeup of the portrayed world. The *host* is the place from which the observer makes his observations. This is ultimately not an entity, but rather a spatiotemporal location which just happens frequently to be identical to that of an entity, and the observer's own body is of course the default choice here. However, in addition to other people's bodies, other locations and times (hypothetical or real), as well as irrational objects are also eligible as hosts.

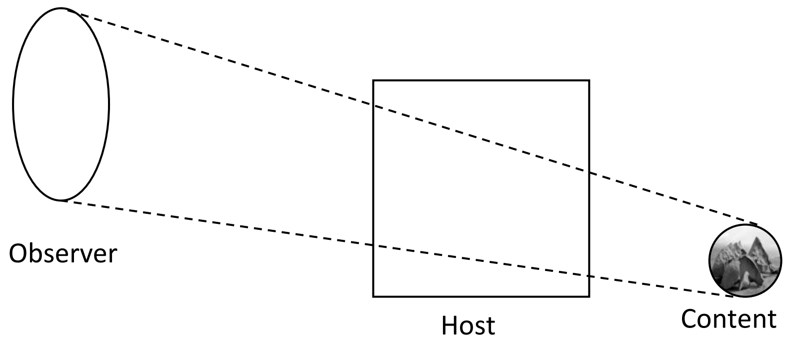


Fig. 5.2. The triad of observer, host, and content. The observer is a judging, rational subject, and the host is a deictic centre that provides the window through which the content is seen. I refer to this window (and its relation to the observer) quite simply as the *perspective*. It has a certain size and a position in relation to the portrayed entity in the spatial, temporal and classificatory dimensions. Unfortunately, it is not possible to visualize this multi-dimensional relationship.

The observer and the host are the building blocks of perspectives, something which is perhaps especially clear in connection with counterfactual expressions. One may observe the world as seen from a certain object as if perceived by a rational being (cf. *If you were that tree*), or an observer may be located in someone else's body (cf. *If you were me* and *If he had been her*). However, it is only normally possible to move a rational observer to the location of an irrational host and not vice versa. This explains why the following sentences are usually thought of as semantically equivalent (Alfort, 2009).

5.1	If that tree had been him, I would not have felled it/him.
5.2	If he had been that tree, I would not have felled him/it.

Both (or all four) of these sentences are paraphrasable as something akin to *IF HIS SOUL$_{OBSERVER}$ HAD BEEN IN THAT TREE$_{HOST}$…*. Taking *he* as the host of the tree's soul is normally out of the question, but this is attributable to an extralinguistic supposition and is not in principle an unbreakable rule. The interpretation that the soul of the tree is the observer taking up residence in his body is not unthinkable; it is simply far less likely than the opposite scenario. We therefore usually take ex. 5.1 and 5.2 to be synonymous. This illustrates the fundamental difference between observer and host as being a rational being and the location of a "window" whence a situation is perceived.

Clear examples of explicit references to the perspective chosen are descriptions of coming and going, or buying and selling, as well as spatial expressions such as *in front of*, and temporal specifications. The metaphor of a window is intended to convey that perspectives include a scale on which the scene and its participating entities are observed; it is possible to zoom in and out on anything portrayed in any dimension by adjusting the size of the window.

Contrary to popular belief and everyday impressions, it is not even self-evident what qualifies as an individual entity and what does not (Jackendoff, 1983); people inevitably impose classification on the world and individualize its components as they see fit, depending on the level of granularity applied (Grenon & Smith, 2004). Thus, a flock of sheep becomes an individual when observed from afar, yet dissolves when single sheep are sought[109]. When a vet approaches an animal, however, it ceases to be an individual and becomes a collective of individual parts. Individuals do not even need to be physical objects (Bloom & Kelemen, 1995); evidence from psycholinguistic experiments shows that “people will consider some aspect of reality as an individual if they are able to predicate important properties to that aspect which do not also apply to arbitrary portions of that aspect” (Wisniewski, Lamb, & Middleton, 2003, p. 587). People may of course refer to either individuals or collectives or parts, though such expressions tend to function differently. Thus, they decide what to consider an individual and then apply an expression at either the same or a different level; one may speak of a flock consisting of individual sheep or of sheep making up a flock. The point to notice here is that when someone speaks of a flock of sheep, it is that person who defines the sheep as being the individuals, and if you see two flocks of sheep approach one another on a distant hillside the flocks are likely to be the individuals in your mind while the sheep are merely their indistinct parts that you know from experience to be individual objects but do not picture as such in that moment[110]. The term that you choose for representing the sheep is dependent on your conception of them as a flock or individual sheep or as a collective of sheep-parts, yet your mental representation of the sheep depends in part on their physical appearance and your perspective on them, but also on your free will and creative powers. As soon as you make the choice, however, you are bound by it, and the sheep are for all intents and purposes to be considered in that way until a new choice is explicitly made (or the topic changes). How you choose to describe the sheep may thus tell the addressee much about the way you conceptualize them and what potential situations you might be thinking that they could potentially take part in.

[109] This example is discussed in various sources (Langacker, 2000, p. 228; Croft & Cruse, 2004, p. 51f; Langacker, 2006).

[110] Of course, it is possible – even likely – that we may have more than one picture of the same scene mentally represented simultaneously (Prinz, 2002).

To take an example, the Danish word *lam* can mean either an individual lamb or an uncountable mass of lamb's meat, i.e. mutton. Knowing whether a user searching for *lam* means one or the other is essential for a successful search, yet in this instance the LAT simply cannot know what was meant. In English, the difference has been lexicalized, and this helps the system identify the meaning immediately; lambs and mutton enter into completely different potential situations, and hence part of the use to which the desired resources will be put is already conveyed.

One case where perspectives are really important is in connection with illnesses. It is well known that people often search for information on diagnoses on the web, and also that this is not necessarily a good idea, because the information that can be found tends to upset patients and relatives unnecessarily. There is a huge difference between what information patients, relatives, professional personnel, researchers and laymen from the general public benefit from. For this reason, the homepage of the Danish Cancer Society (www.cancer.dk) that services all of these user groups is carefully constructed so that each type of user is shown appropriate information presented in a suitable manner (though all users are free to select another user group if they so wish – an important ethical precaution). Whether a query consisting of the name of an ailment refers to the user's own health in the past or present or in the potential future or to someone else, makes a significant difference to the relevance of various types of information, and not least the presentation of it.

Technically, according to the definitions adopted here, it is the person for whom the information is relevant that is the *user*, not the individual that is carrying out the search. Because a user is distinct between searches, each user is associated with a distinct set of spatiotemporal coordinates. This is exploited in contextual IR, especially as regards mobile platforms. In this way, users are hopefully given results that are relevant in their spatiotemporal context. However, the user's coordinates need not refer to the actual individual doing the searching, because a speaker is always free to portray a situation from another perspective. Obviously, some are more appropriate than others in a given situation. For instance, in describing a scene as it occurs in front of the speech act participants, it is natural to portray it from the perspective of the speaker. On the other hand, there may be very good reasons for deferring a portrayal to another perspective, either spatially (as viewed from another angle), or temporally (as experienced from another point in time), or even as seen from a hypothetical situation that differs slightly (or distinctly) from the real world. Accordingly, a query such as *restaurant Valencia* does not necessarily mean that the user is currently in Valencia, or even wishes to go there. It may refer to a potential future state of affairs, or it may simply be a request on someone else's behalf. Similarly, a search for *queen Buckingham* need not refer to the present queen, since the intended time may not be the present, or simply generic so that information on any and all (relevant) queens connected with Buckingham would be considered relevant.

What this means is that collecting information on the actual searcher is not necessarily helpful (except as regards psychological characteristics such as level of uncertainty and expertise). Rather, the perspective from which the references in the query are constructed should ideally be identified. The actual searcher's time and place and worldview are a good guess when no other clues are available, but it is far from certain that this proxy represents the correct interpretation.

Communication invariably involves the assumption that the interlocutors share perspectives to a sufficient degree to make a reasonably correct interpretation feasible, and it follows that every word must have a perspective assigned to it, though this need not necessarily be specified explicitly. However, until an expression is produced, the speaker's mental world does not contain any specified perspective; rather, any point of view may be adopted with every word uttered about it – in fact, one word may contain two perspectives simultaneously, as predicate reference and term reference may be applied from different perspectives – something which is simply not possible in the physical world (Alfort, 2009). The following classic example shows clearly how predicate and term reference can be applied from mutually incompatible perspectives on the same term.

5.3	Oedipus wanted to marry his mother.

In this case, Oedipus did not know that his intended bride was his mother, so the predicate reference relating to the predicate *mother* must clearly be considered as applied from the point of view of the interlocutors, by activating shared knowledge. However, the term reference is made from Oedipus' point of view (as well), as he could of course identify the one he wanted to marry as an individual. The host is moved back and forth in space and time (and reality) through a narrative in this way, and in some cases even the observer may change. This can be traced in texts thanks to linguistic clues and by sheer logical inference, but the question is how the perspectives of query terms can be discovered.

5.2 Bisubjectivity

When I sat down to write my master's thesis on referential phenomena, my basic *research-for-understanding* standpoint forced me to pose myself some very deep questions: What is this thing that we call *language*, and how does it really work? I came up with a fundamental theory of reference that revolves around what I have come to call the theory of **bisubjectivity**. The principle of bisubjectivity has profound consequences for the way we use language in all of its shades, and as search queries are clearly instances of language use, bisubjectivity is at the very heart of the present thesis, right from a user's phrasing of a query over the assessment of the relevance of a resource to the evaluation measures employed for calculating search engine effectiveness. The phenomenon was not unknown; in Relevance Theory it is currently known in a specialized function under the slightly more cumbersome name of *mutual parallel adjustment* (Wilson & Carston, 2007). However, I wish to stress that this is a highly general phenomenon that goes far beyond lexical adjustment in language; we shall see later, for instance, that it enters into relevance assessment of resources in search.

5.2.1 The double act of referring

It is an old enigma of linguistics that language is both independent of and constrained by the world that it portrays; speakers are free to say anything they wish in principle, but some utterances are still inappropriate in a given context (Lyons, 1977, p. 577). For instance, in languages with obligatory gender agreement, using an inappropriate gender form is simply wrong. This is so even if the word may be used with both genders, if at the same time it is clear which gender the referent has. In most cases, the following sentence cannot, therefore, be considered correctly expressed if said by a man, in spite of being formally entirely well-formed and perfectly acceptable if spoken by a woman[111].

5.4	Je suis heureuse. 'I am happy$_{fem}$.'

The bisubjectivity of language is responsible for this phenomenon; the speaker constructs a mental world and is then bound by it in forming a truthful utterance, and hence the sentence can only be used if he pictures himself as a woman (at the moment of speech). The alternative is an uncooperative attitude that will make it impossible for hearers to interpret his messages.

The fact that reference is made to mental images rather than real-world entities means that we can structure our subjective model of the world as we please and adopt different viewpoints of the same scene, both as regards the physical and social relations between individual constituents and our lexical descriptions of them. Here is one example sentence borrowed from Talmy (1988).

5.5	There is a house every now and then through the valley.

[111] Except, as pointed out by Lyons, by a man playing a woman's part in a play, etc.

In this sentence, a special physical perspective is imposed on the scene in which the subject is moving through the valley as if with a mobile film camera (Langacker, 2000, p. 212). It is not likely that the speaker is actually carrying out this movement as the sentence is uttered, so it is an entirely constructed description based on a dynamic mental picture of the world that the speaker wants to portray. Here is another kind of example from my own master's thesis.

5.6	Being such a hero you might just help me with this.

In this sentence, the speaker does not in fact consider his interlocutor a hero. The sentence is ironic and the term *hero* is actually ascribed from the hearer's point of view, in contradistinction to the rest of the sentence which is constructed from the speaker's point of view (Alfort, 2009). Shifting the point of view of lexical choice in this way in the course of a single sentence is much more common than one might initially think. Some perspectives are entirely hypothetical and induce changes in the world portrayed by them.

5.7	If they had built that house a bit further off, we might have been able to see the lake from here.

In this case, a hypothetical perspective is adopted in which the elements are not placed in the same way in relation to each other as in the portrayed real-world situation. This is no problem at all because of the fact that reference is never made to the real world but only to the mentally constructed scene. We can only observe "hypothetical worlds" by applying hypothetical perspectives, because what the sentence refers to is the mental world constructed on the basis of the chosen perspective, which is not in any way required to correspond to anything that actually exists in physical reality. Of course, in order to be intelligible and informative to each other, we must regularly anchor our speech in reality, but this is a purely pragmatic necessity and not a fact connected to the workings of reference.

According to the fundamental principle of bisubjectivity, we are free to choose, yet are bound by our choices, or in other words, there are laws that we must conform to, but at the same time we are the ones who write the paragraphs of those laws. Crucially, this is true for all linguistic acts. The sentence in ex. 5.7 cannot be claimed to be an incorrect representation of the world – even though it does contradict the actual state of affairs – because it is only supposed to represent the mentally constructed hypothetical world in the speaker's mind, and this it succeeds in doing. This makes it really hard to study a phenomenon such as language which is subject to bisubjectivity.

> Exactly why is it that it would be ungrammatical for a speaker referring to a woman but thinking that it is a man to use the feminine gender?

From the analyst's point of view this situation simply involves a woman and a speaker seemingly referring to her as a man. The tricky question is whether this mismatch is caused by the speaker having a different definition of what qualifies as a WOMAN, or whether he is simply mistaken about her gender (or wishes to mislead the hearer). The latter case is a non-linguistic phenomenon that shall not concern us here. In the former, the use of a masculine expression is dependent on characteristics of the referent as well as on the conceptual structure of the speaker's mind; it cannot be ascribed to one phenomenon alone. This is something that is of great importance in the assessment of needs and relevance – two topics which are vital in information retrieval research. Documents evaluated by a user may be thought relevant or irrelevant depending on features of those documents, but judged against criteria defined subjectively by the user, and so it is not always possible to decide whether they are relevant because of their own desirable characteristics or because of the user's relatively slack criteria. Thus, the same basic principles govern search and language production, not only because search involves language, but because both involve human assessment and classification.

5.2.2 Bisubjectivity and cybernetics

All living beings are capable of self-adjustment and consequently constitute **cybernetic systems** (Heylighen & Joslyn, 2001). Through a cybernetic circle, the system evaluates its current state, performs changes as necessary and ends up in a new stable state. This phenomenon permeates human behaviour and ought consequently to be extremely important in studies of search (Brier, 1996). Nevertheless, researchers have tended to mistrust cybernetic principles as an approach to IR (Spink & Saracevic, 1998). This pessimism seems to be unfounded and based on a lack of mutual understanding between cyberneticists and IR researchers; for example, the latter consider *feedback* to be what the user returns to the system by clicking hyperlinks and opening documents, whereas to a cyberneticist *feedback* occurs in the form of responses from the system and the cybernetic process involved is one initiated by the human agent. This appears to cause terminological misunderstandings (Spink & Saracevic, 1998).

Relevance judgment is a cybernetic system just like other human behaviour; it is just that the most basic model of cybernetic circles has to be tweaked slightly in order to accommodate this kind of behaviour. The simple model of a cybernetic circle is one that can be used of a central heating system, for instance (fig. 5.3). It includes a *sensor* measuring the actual temperature in the room, a *controller* comparing the outcome of this analysis to the "*set point*", i.e. the temperature aimed for, and an *effector* that warms up the room when the temperature has been deemed to have fallen too low. The problem with this simple model is that the *set point* is treated as though it were a simple value, which is natural given this heating example, but in fact the set point is a model (very simple in this case) consisting of a desired value and a range of acceptable deviation from this standard. The controller needs to know when a temperature as measured is sufficiently far removed from the set point (in the narrow sense of the word) for the effector to be switched on.

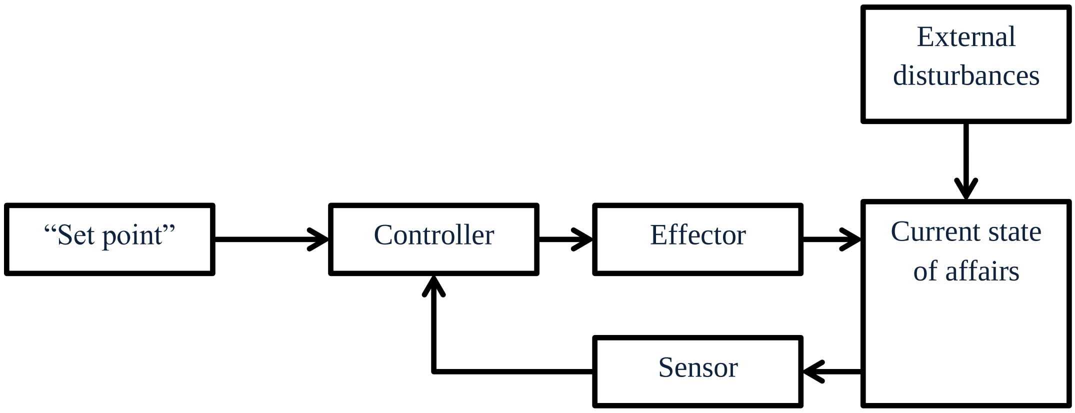


Fig. 5.3. A simple cybernetic circle as traditionally depicted (Ivanovas, 2010). What is usually called the *set point* is more properly a model.

In fact, this model may even be said to be couched in a context, which allows considerations such as whether the desired temperature is fixed or varies with external conditions. It might be for instance, that the time of day was significant in that respect, or even the current whereabouts and activities of the inhabitants. A simple unintelligent heating system is unaware of such contextual clues, but some are not, and when it comes to human behaviour, the input should always be thought of as a model couched in a context. Furthermore, humans are able to change the model at will. In fact, they can hardly refrain from doing so, as context changes continually and this immediately affects the model. Thus, the cybernetic circle that we need is slightly more complex (fig. 5.4). What makes human behaviour bisubjective is the fact that the same sensor – the human mind – evaluates the current state of affairs as well as the suitability of the current model in context, and is hence not only able to change the state of affairs but also the model that prescribes when and how to change it. This is true of both language use and relevance judgment.

In language production, the *model* is the mental world constructed by the user in order to be portrayed. This model can be modified during the production of an utterance (though probably most often before the actual uttering of it) by applying certain perspectives suitable to the conveyance of a certain message, or even in more invasive ways by altering its composition. For instance, it is possible to suppress certain referents that the speaker does not want the hearer to know about. The real world at most enters the circle in the form of *context*, whereas the *current state of affairs* represents various potential expressions which fit the model more or less perfectly and portray the situation in a way that suits the speaker's intentions more or less well. The circle goes on until the speaker is satisfied that the currently evaluated potential expression is optimal or at least sufficiently well-formed to be used. Most of this happens subconsciously, so there is not usually any actual comparison of expressions involved.

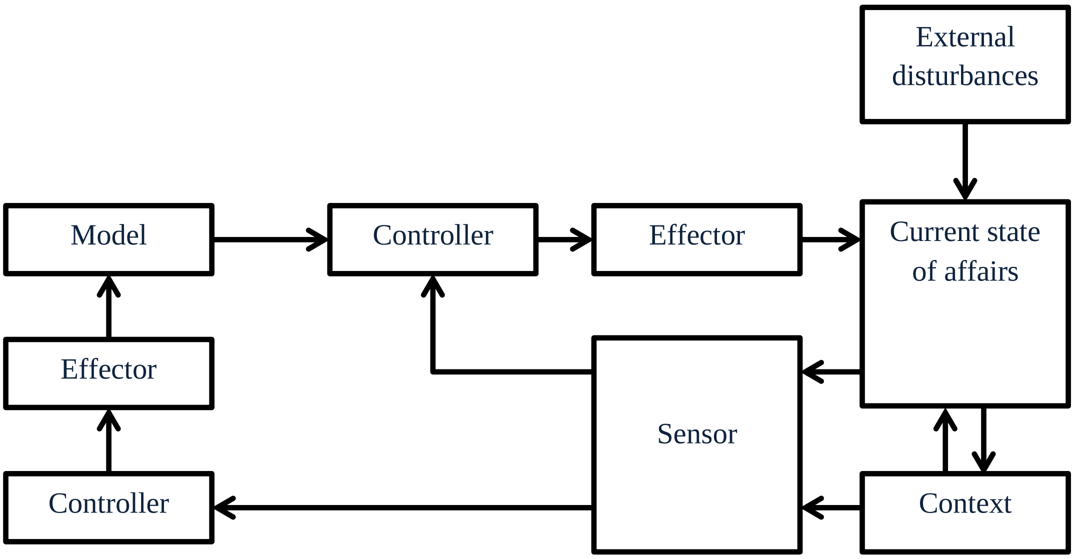


Fig. 5.4. A model of a cybernetic circle adapted to human behaviour.

Relevance is often thought of in the same way as heating systems, i.e. as involving a set point to be reached. However, relevance is *not* a matter of values but rather of suitability to certain contexts, just as with language production. The *model* (replacing the *set point* in the simple version) in this case is likely to be some kind of ideal of resources which would be relevant in the circumstances. It involves not one but several ranges of acceptable deviations from the perfect resource along a set of dimensions, and is strongly dependent on context and hence subject to constant revision by the human sensor.

The effect of this bisubjective cybernetic process of creation is that given an output it can be very hard to reconstruct the component parts of the input, and also that the sensor itself – the human being involved – may have severe difficulties tracing the development of an output. It may not be simple, for instance, for users to specify exactly why they ended up judging a particular resource potentially relevant or likely to be irrelevant, as the reading of the resource itself may have affected the user's idea of what was needed in the first place, hence providing an amended model of what the ideal relevant resource would be like.

Users assess the relevance of documents according to certain – usually unstated – criteria, but at the same time they are the ones setting the criteria, and for this reason it can be very difficult to determine exactly why somebody rated a given document as being relevant or irrelevant, given that it may be due either to a feature of the document or to a criterion as applied in that instance. A document may even be judged according to both "points of subjectivity" simultaneously, i.e. a user may acknowledge that the resource has some relevance according to certain analogue scales, but at the same time decide that it does not surpass an ultimate digital threshold (supplied by the user) that will allow it to be categorized with other relevant documents, and thus – absolutely speaking – it must be classed with the non-relevant ones.

A cybernetic system may regulate itself by *feedforward* and *feedback* (Heylighen & Joslyn, 2001). A feedforward mechanism warns the system of upcoming external changes in the environment which necessitate certain actions in order to retain equilibrium or move towards the goal of that system, whereas a feedback mechanism takes stock of the actual situation and measures any errors or deviations from the expected or desired state which might occasion similar actions. In search, different users[112] are sure to prefer different ratios of feedforward and feedback; some expend their energy on assessing resources before accessing them properly in an attempt to predict relevance, while others grab more or less whatever is suggested and evaluate the relevance of the actual resources via feedback. Feedforward and feedback may not provide compatible information. For instance, a user may decide that a paper has a very interesting title and thus looks promising, while a quick perusal of the beginning of the actual document soon reveals that it is actually likely to be extremely uninteresting. In this circumstance, the cognitive feedback mechanism of relevance assessment is likely to produce a warning that this material is probably not worth attending to and ought to be rejected.

5.2.3 Digitalizing an analogue world

While the world is notoriously **analogue** and infinitely detailed, it is a basic limitation of being an observer that we can only conceptualize and (especially) portray objects in a **digital** manner, as pointed out by Gershenson and Heylighen (2005). The authors follow Gershenson (2002) in distinguishing the *absolute being* from the *relative being*.

[112] Note that this still refers to the definition of *user* adopted here, i.e., every search has a new user even though it is the same individual doing the searching.

> The absolute being (abs-being) refers to what the thing actually is (Kant's Ding-an-sich). The relative being (rel-being) refers to the properties of the thing as distinguished by an observer within a context. Since the observer is finite and cannot gather complete information, rel-beings are limited, whereas abs-beings have an unlimited number of features. Therefore, there exists an infinity of potential rel-beings for any abs-being. (Gershenson & Heylighen, 2005, p. 50)

The constructed world in our minds is "pseudo-analogue", i.e. it represents our perception of the real world (if there is one, i.e. to the extent that the mental world is not fictive) relatively truthfully and only regularizes whenever this is necessary in order to fit into the boxes provided by our cognitive categories and sensory limitations. A linguistic expression based on it, on the other hand, is "pseudo-digital", i.e. it basically classifies everything according to the terms chosen and does not allow for much vagueness, except by using superordinate, less specific terms. It does to some extent allow grading, but only in a limited number of dimensions – not according to a sheep's degree of *sheepness*, for instance – in language, something is either a sheep or not a sheep. Hearers must convert this pseudo-digital signal into a pseudo-analogue world of their own.

It will be clear by now that there are two points of subjectivity in our process of expressing ourselves; the first is the construction of the pseudo-analogue mental world, and the second is the pseudo-digital expression of our portrayal of it. Thus, we may claim that something is a sheep and yet be of the opinion that it is in fact only a sheep *to a certain extent*, or *for a given value of 'sheep'*, but we will have difficulty expressing it. In other words, we might decide that the referent in question was a sheep to a certain extent on an analogue scale while not qualifying as a sheep in a digital sense because it did not pass some *ad hoc* threshold of true sheepness, such as having four legs, not being made out of plastic, or not having reached a certain age. The same is true of relevance assessment in search, which has an analogue and a digital side to it that function in a bisubjective way, as described in depth in section 6.3.1.

5.3 Homing in on a target – Navigating referential space

Having chosen what to refer to, the user must decide *how* to refer to it. In principle, there is no limit to the number of different predicates that can potentially be used in reference to a given entity. That depends entirely on the situation and the intention of the speaker. This section elaborates further on the almost unlimited amount of subjectivity in the expression of a query.

Lexical variation basically covers two phenomena: The choice of **scale** and what I shall call **qualitative aspect**[113]. Sometimes people search for highly precise concepts, and at other times they search broadly. This represents a narrow and a broad *scale*, respectively. Very broad categories may probably often be seen as a kind of *navigational* search; such users are really performing what we shall in Part III call *anticipatory header* searches, i.e. they are looking for a page with the title in question. These searches stand out clearly in search logs from supermarket e-shops. When users search for *fish* or *sweets* it is hardly the case that they do not care what kind of fish or sweets they get. They simply want to be taken to the virtual department in question in order to examine the alternatives and decide on a specific product. It is conceivable that the difference between broad and narrow reference could be exploited in a LAT since it is clearly correlated with a difference in intention. Often, users issuing very broad searches would be served better by taking them to an overview page rather than attempting to deliver relevant search results. A user searching for *animals* in a pet shop site would be likely to be after a list of the animals available in the shop. However, it would probably be easier for the user to choose between the animals via the onsite navigation structure rather than by clicking through many pages of search results relating to animals. Much could thus be done to optimize results presentation whenever semantic extensions of this type are performed.

The other parameter of variation is *qualitative aspect*. While *scale* covers differences in generality or specificity, the ontologists' term *aspect* deals with the characteristics of an entity that the speaker chooses to focus on. Two examples will suffice to illustrate this point: A rat may be called either a *pet* or a *pest*, and a pound of clay may be considered a piece of art or a lump of material. This gives further emphasis to the observation that we create our world while speaking rather than portraying a ready-made one. As with scale, what qualitative aspect users choose is of vital importance when it comes to understanding their intentions. Luckily, it is also relatively easy to decode, as it is expressed directly in the choice of predicate. The challenge is to relate the various qualitative aspects that are *colocalized* in the same perceived entity to each other in reasonable ways so that documents on *rats* are returned when users search for *pets* if and only if they really do deal with rats *as pets* and not as pests.

The use of broad terms may cause problems in search, because it is not always given that the user is also interested in information on subordinate concepts. For instance, in a pet shop it may be natural to expand a search from the broad term PET to CAT, DOG, GUINEA PIG and RAT, but if the same search were performed in the site of the local authorities this ontology would be unfortunate, given that documents

[113] Among ontologists this phenomenon is normally known simply as *aspect*, but as this is a term that linguists use when talking of a different phenomenon that I shall also be introducing here, we shall have to use the fuller description *qualitative aspect* when talking of this classificatory kind.

on rats would be more likely to concern pest control than the regulations concerning pets that the user might have been looking for. This is an example from the actual implementation of a search engine for the Danish authorities, and here it was furthermore found that guinea pigs likewise had to be removed, because by a quirk of history the words for GUINEA PIG and PORPOISE happen to be homophonous in Danish ('marsvin'), and the authorities tended to deal with porpoises rather than guinea pigs. Only some concepts should be expanded into their subordinate parts, but exactly when this would be beneficial is a matter for further research.

The Danish term *marsvin* has to be narrowed down to one of its senses. However, Wilson and Sperber (2004) demonstrate that a term such as *bank* may not only need to be narrowed in sense from FINANCIAL INSTITUTION OR RIVER BANK to FINANCIAL INSTITUTION, but also loosened in sense from BANK to BANK OR CASH DISPENSER in a context such as *I forgot to go to the bank*. In a similar manner, a query term may need both **narrowing** and **loosening** according to context. Imagine a user entering the word *bank* in a context-aware mobile search application; probably, cash dispensers would be at least as satisfactory in most everyday cases. The only way to perform such loosening correctly is by understanding why the user needs the item in question – cf. an erroneous loosening from BANK to BANK OR OTHER PUBLIC BUILDING. Such a loosening might easily come to the computer's "mind" after having consulted an ontology, yet it would rarely be an appropriate inference. Only by understanding that what the user wants with the bank is to make a withdrawal is it possible to make an appropriate loosening. Narrowing and loosening represent one of the reasons why we may need users to describe their task properly before attempting to interpret the query; there is so much which is rarely expressed explicitly in queries.

5.3.1 The classificatory dimension

Scale and qualitative aspect both deal with the type of an entity referred to, but in different ways. Alfort (2010) demonstrated that they may be considered dimensions in the manner of space and time; in fact, scale and qualitative aspect could be said to represent the vertical and the horizontal dimension of classification, respectively. The former is about choosing a more general or specific term, and the latter about choosing a different parallel term that describes the object equally well from a certain point of view[114]. The temporal dimension is replete with metaphorical expressions and conceptions like *a short time*, *time moves*, *we are approaching the future*, *from the point of view of fifty years ago*, and so

[114] The following is a synthesis of the important points from Alfort (2010).

on, and we have long ago taken advantage of the fact that language treats time as a dimension parallel to space and have learned a lot from the recognition of this metaphor[115].

> Commonsense distinguishes between *things* (*spatial objects* like houses and computers) and *events* (*temporal objects* like bank transfers and computer repairs). In the wake of relativity theory, however, time is viewed as another dimension of objects on a par with the traditional spatial dimensions. Considering the consequences of this scientific theory (or theories), some philosophers and computer scientists have come to believe that the commonsense distinction between things that are and things that happen should be abandoned in favor of a unified viewpoint. According to these revisionist researchers, *everything* extends in space *and* time, and the distinction between things and events is an (ontologically irrelevant) historical and cognitive accident. (Masolo, Borgo, Gangemi, Guarino, & Oltramari, 2003, p. 7; emphasis original)

This is once again a reminder that there is no difference in kind between nominal and verbal reference; events and entities both extend in time and space in similar manners. We may now continue the development mentioned by Masolo *et al.* by recognizing that what appears as points or regions in space and as periods of time correspond to the sets of potential referents of classificatory categories; very general categories have many potential referents and correspond to wide regions, whereas very specific categories have fewer potential referents and correspond to smaller areas. Not only does this place us in the happy position of being able to draw on the well-known metaphor of perspective in order to understand classificatory phenomena better; it also seems that speakers do treat classification in terms of this metaphor. We say things like *from the point of view of forestry*, in which case *forestry* is not to be understood as an entity but rather as a topic or mindset on which all subsequent classification is based (if in any way relevant to forestry). We also speak of *broad topics*, of *closing in on a topic*, and so on; statements which testify to the fact that we tend to think of classification as something navigable – a plane of sorts. This allows us to analyse classificatory choices with tools developed for the other dimensions. Just as a speaker can zoom in on an entity in the spatial and temporal dimensions when applying a perspective, this is equally possible in the classificatory dimension, resulting in a more detailed description. If we have a detailed knowledge and interest in a restricted region, then we generally employ a narrow scale with a high level of detail, we explicate spatial and temporal details in our portrayals of a situation and we use relatively specific categories in our classification of entities and actions. If, on the other hand, we intend to make a general statement, then we zoom out and let the

[115] The scientific question whether time really is a fourth dimension is irrelevant in this connection.

internal distinctions become blurred in space and time as well as in classification and the type of the entity in question becomes a mere point within a larger region[116].

5.3.2 The topography of the various dimensions

A dimension can be topographically characterized by stating whether an entity is a point without internal structure (from the point of view adopted) or a region with diversity. This question shall be referred to as **aspect** in the proper linguistic sense (the co-existence of the two meanings of aspect is unfortunate and confusing, but the situation would hardly be improved by introducing additional terms). It can also be located relative to some reference point, typically associated with the speech situation. In the temporal dimension, this is known as **tense**, but as that term is so obviously restricted to time I shall choose the cross-dimensional term **relativization**. Also, regions may be distributed across the dimension, so that one reference may do for several referents. I shall call this **distribution**. Lastly, points may be **internally ordered** as with contiguous points in space or consecutive moments in time. We shall explore each notion in turn.

- **Aspect: Point vs. region**

  In time as in space, a referent may cover a region or be considered a mere point. The latter depends on the chosen scale; a point always turns out to be a region if the observer zooms in on it, and *vice versa*. Even in the classificatory dimension(s) there are points and regions. A classificatory point is a reference to a specific type with no internal variation. This is only normally possible if there is no more than one individual answering the description, because there is always some variation between individuals. However, as with spatial regions that become points when seen from afar, classificatory regions also lose their internal structure when one zooms out so that a class may be considered completely homogenous because *irrelevant variation is ignored*. This is less natural with very general terms than with more specific ones as more zooming and consequent blurring is needed. A classificatory region corresponds to a heterogenous category.

[116] Among ontology modellers, this has resulted in a difference of opinion between those who want to portray the world in the way they think is most scientifically correct (the *revisionists*) and those who want the ontology to mirror our cognitive and linguistic viewpoint (the *descriptionists*) and therefore keep the spatial and temporal dimensions apart (Masolo, Borgo, Gangemi, Guarino, & Oltramari, 2003). I would like to emphasize that what is proposed in Alfort (2010) is a cognitively based ontology, even though I go one step further in bringing classification in line with the other dimensions. I certainly do not claim that the classificatory dimension has physical reality; it is a purely cognitive notion. However, I do believe that both the classificatory and the temporal dimensions are to some extent treated in parallel with the spatial ones cognitively, and that this assumption does not necessarily make an ontology “revisionist”; treating different kinds of dimensions in parallel is very different from completely ignoring the distinctions.

● **Relativization: Location of region**

Tense locates the referent in relation to some point of departure, which is only recoverable from the location of the host (Comrie, 1985). Such relativization is commonly encountered with space as well, though this is less often remarked upon because it seems so natural to us to locate things relative to the spatial coordinates of the host. Finally, relativization occurs in the classificatory dimension too – with subjective predicates that require a tacit standard of comparison, such as *she is tall* or *it is heavy*.

● **Distribution: Extent of region**

In the spatiotemporal dimensions, simple distribution in the guise of a non-singular reference restricts the reference to several entities describable by the same predicate. Such applicants are spatiotemporally distinct, but classificatorily identical. This means that you cannot rename one entity without having to rename all the others as they are by definition classificatorily identical from the adopted point of view. Distribution in the classificatory dimension refers to several types, which are classificatorily distinct but spatiotemporally identical; this is known as **colocalization**. It entails that you cannot move a spatiotemporal entity (i.e. assign it a new position in space and time) described by one predicate without moving the extension of the other type(s) correspondingly. In other words, moving a sculpture entails moving the clay and *vice versa*.

● **Internal ordering**

The last type of topographic information is *internal ordering*. In space and time, some points are obviously contiguous and are capable of being mapped in relation to each other. This is not so clear in classification, though some categories are certainly more similar than others. In space, we order points along three dimensions. In time, along one. In classification, there is an unknown, possibly even infinite, number of dimensions, and this means that ordering classes for similarity is not immediately possible.

Worldly phenomena are not static, and descriptions of them consequently only apply at certain points in time. A speaker uses a certain predicate in reference to a given object, and specifies to some degree when and where it is meant to apply. On the accommodation layer, it does not necessarily really apply in all these cases, because the speaker may have generalized somewhat by excluding irrelevant details. Just as the predicate *apple* in *He eats an apple every day* has a temporally distributed reference rather than referring to a single individual across time and space, the habitual predicate *walks* has a distributed reference in *He walks to school*; it is applied as a verb of general validity that is effective at any time within a relevant time span, but it only effectively applies to applicants in the subjective world at those times when the speaker is on his way to school. The interpretation of the predicate's reference thus accommodates to the world by having a temporally distributed interpretation enforced. This often has morphosyntactic consequences in that aspectual forms (such as the habitual) are ascribed in accordance with the accommodation result.

It is not immediately clear what to do about this extremely complex situation. Language is immensely difficult to deal with in a computational environment due to the inherent bisubjectivity which means that all of these topographical details of the various dimensions involved are on the one hand chosen by the user and on the other accommodate to the features of portrayed phenomena. Consequently, it can be really difficult to disentangle what belongs where in interpretation. If a user chooses to use a general term rather than a specific one, does this mean that the user wishes to refer to a general phenomenon or merely to portray a specific phenomenon in a general way due to a wide scale? Would a user searching for *illness* be interested in material on *diabetes*, or only in material which relates to illness in general? These are the kind of serious ambiguities which we should be worrying about at present rather than mere lexical homography. In full discourse, much of this ambiguity is resolved in context, but the context needed is exactly the kind that is typically missing from queries, such as expressions of opinion and judgment, and conversational particles. There is no easy solution to this problem, but if one were to be found it would be necessary to evaluate the effectiveness of such a system, and for this purpose a method for evaluating bisubjective phenomena is developed in section 5.4.3. Much further research is needed on these issues.

5.3.3 Colocalization

It is an old discussion whether two entities can possibly be spatiotemporally colocalized, or whether they are by definition facets of the same entity.

> Are the statue of Apollo and the clay from which it is shaped different entities or one and the same?

One opinion is that the two classes ascribed to the statue simply refer to different classificatory *aspects* of the same spatiotemporal entity. Which one you choose will depend on your view of it, but also the use to which you intend to put the entity. Reductionists prefer seeing colocalized entities as different *views* of the same spatiotemporal object (Masolo, Borgo, Gangemi, Guarino, & Oltramari, 2003). This shows us the spatial metaphor in use once again; just as an entity can be observed from various spatial or temporal viewpoints, it can also be considered from different classificatory viewpoints. If classification is accepted as another dimension, then any thought of true colocalization of different entities must be rejected; if two things are identical in type as well as in space and time, then they are one. On the other hand, if they are only identical in time and space but not in the classificatory dimension (i.e., a classic instance of colocalization) then they are colocalized in the spatiotemporal dimensions only, yet the reductionists' claim is satisfied too, in the form of a metaphor; there is spatiotemporal colocalization and the difference is a question of viewpoint – or perspective – in the classificatory dimension(s). The two standpoints are thus effectively merged into one.

If classificatory differences are translated into locations in the classificatory dimensions then different instances of the same type are cases of colocalization of distinct spatio-temporal entities in the same region of the classificatory dimension (note that it is usually a region and not a point, because instances will normally differ among themselves to some extent). Colocalization of several temporo-classificatory entities within the same space may initially sound strange as different types do not count as one entity even if they occur in the same spot at different times, but in fact everything that changes, including human beings, is an instance of the same spatial region (in a relative sense) moving and changing in time and type. Time and class must be correlated, and they must be continuous regions, just as the sculpture of Venus must be situated in one spatiotemporal region where time and place is correlated. Even colocalization of various spatio-classificatory entities within the same time but with a spatial and classificatory extent occurs when the same individual changes its character over space, which is of course entirely commonplace. This does not in any way hinder us from considering it a single individual. Again, space and class must be correlated, changing smoothly over the spatial region.

With the traditional spatio-temporal colocalization, we can refer to the different qualitative aspects of the spatio-temporal entity by using different terms. The same person may thus call the same thing by different names depending on the intent. With temporo-classificatory colocalization, we only have names for the various phases in certain cases. Mostly, reference to one temporo-classificatory entity can only be made by use of modification. With spatio-classificatory colocalization, we sometimes have names for different regions of an individual and otherwise we use modification.

These matters are in fact far more important for our purposes than one might initially think. For instance, a user searching for *London 1835* explicitly selects one temporal aspect (1835) of the spatio-classificatory entity that we term London. The different times are colocalized within that spatio-classificatory region, and any time can be chosen. It is of course equally possible to zoom in or out and choose a longer or shorter span of time, or all of the time in which London has existed. Thus, there is a vertical choice of scale and a horizontal choice of qualitative aspect in connection with time as with space and classification.

It is clear that modifiers play an important role in the expression of qualitative aspect (and scale) choice. It may be asked how the modifier *1835* differs from a modifier that signals qualitative aspect or scale. Why, for instance, is *London 1835* materially different from *London hairdressers*? The answer is that London 1835 is an entirely different entity from London today. There are few similarities between the two apart from the spatial distribution (and even that differs greatly at the margins) and the unchanged name *London*. The classificatory dimension is thus distributed and very heterogeneous across time and space. With *London hairdressers*, on the other hand, even if we do construct this as a modification of *London* rather than the other way round, the user is simply after information on London (in an unspecified relevant spatiotemporal aspect of it, most probably at the approximate time of the search and with a scale such that the expression covers most of the city, though some parts of it are probably more relevant than others) as regards (the distribution of, etc.) hairdressers. The reference of the term *London* is unaltered by the addition of the modifier, which only modifies the *topic*. This difference between modifying the concept and the topic behind one and the same label is something that we shall return to in section 9.2.1 in connection with query analysis.

5.4 Dealing with perspective

Bisubjectivity is a feature of all kinds of communication, and – being linguistic expressions – queries work in the same way. Thus, just as an utterance does not describe the world directly but rather the mental model that the speaker has created of it prior to producing the utterance the query does not describe a need directly but may refer to the contents of a hypothetical resource which would be ideal for satisfying it, the situation that caused it to arise, or an action that would help satisfy it, etc. At times, the query may in fact not have anything to do with the user's actual information need as such but may simply be an aid for the user to locate the ideal resource. This is clearly the case with so-called *navigational* search, for instance (cf. chapter 7); when issuing a navigational query, the user intends to locate a certain homepage (or rather, a certain location on the internet) and provides some information that is likely to lead to it, such as the name of the organization in question, or even parts of the URL. Contrary to what is usually claimed, the user's need in such cases is probably not just that of getting to a certain page; often, there is some specific information that the user is after and thinks will be available from the page in question[117]. Thus, the query does not represent the need but merely the hypothetical resource that is thought to be able to satisfy it. It is interesting to note that this behavioural phenomenon does not stop at traditional navigational queries. People sometimes search

[117] Sometimes the user wants to be taken to the site or resource in question simply in order to experience it (Rose & Levinson, 2004). Even in these cases, however, it is not the act of bringing the user to the desired virtual location but the contents of the resource that satisfy him or her.

for one topic simply in order to find information on another which is likely to co-occur with it. This is what Teevan *et al.* (2004) call *teleporting queries*. For instance, if a user wanted to know whether Gerald Durrell's wife is a biologist and did not know or remember her name, he might reasonably settle on searching for *Gerald Durrell* in the hope that the information would be available through resources on him. In fact, it is well known that users rarely know exactly what their need is, and it might indeed be said that all such instances are cases of searching for one type of information while actually needing something else; the user is hoping that searching for a certain topic will help him or her identify truly relevant information. In such cases, the hypothetical ideal resource is often very vaguely defined as simply containing information on the topic, or in other words, the resource that the user is imagining (or is able to visualize) is restricted to the representational facet (cf. chaper 2). If users have more precise ideas about the text of the ideal resource, they anticipate the formal or even substantial facets of it. On the other hand, they may not even know what topics to search for and may thus only anticipate the interpersonal facet of the resource searched for, i.e. they may know in what ways the ideal resource would be helpful, but not what its contents would be. In cases of more vague motivations such as a need for entertainment the ideal resource may even be identified through its effective facet alone, i.e. as a resource with certain qualities that will affect the reader in particular ways without being distinctly helpful in a given situation. Anticipating the intentional resource presumably amounts to knowing exactly what resource one is looking for, which is the case in refinding tasks. These different types of search are discussed thoroughly in Part III.

Search is the comparison of two intentions – that of the user, expressed through the query, and that of the author, coded in the document, and the system needs to be able to interpret both what the user intended to communicate to it, and what the author intended to inform readers of. The makeup of the imaginary resource that the user has in mind constrains the way the query may be expressed. If a user who wants to buy a new car settles on *cars*, then that is probably because she imagines an ideal resource with information on various cars, whereas if she chooses the phrase *new car* she is more likely to have envisioned a site with information on how to buy a new car. Consequently, the form of the query can tell us how the user envisages the ideal resource. If the user was describing a problematic situation in need of a solution, such as a situation involving one or more cars, the same morphological differences could tell us how this situation was perceived. In that case, a generic phrase such as *cars* would imply a generalization of events, i.e. the user would probably be looking for information on how to deal with cars in general rather than in a particular situation.

5.4.1 Genericity as an idealized generalisation

As mentioned above, Alfort (2009) showed generics to be a special perspective rather than a characteristic type of reference. It appears to be a hypothetical perspective that covers all relevant instances of something. Exactly what is relevant is left to contextual inference. Thus, the generic topic *sea monsters* refers to the sea monsters present in any and all imaginable (relevant) situations characterized by their comprehending a sea monster. It does not usually refer to the type as such as generally believed. It *can* be used to make such species-level statements or scientific definitions, but that is a specialized use, which is not even limited to generics.

5.8	Sea monsters are a kind of monsters.
5.9	A sea monster is a type of monster.

In most cases, a lot of instances which could legitimately have been included under the label in question are tacitly excluded from such generalizations. For instance, even though it is undeniably true that *owls lay eggs*, there are many owls that do not do this, such as male owls, malnourished owls and owl chicks (Kadmon & Landman, 1993). Data which may be deemed sufficiently irrelevant for this kind of exclusion include not only classificatory but also spatial and temporal instances. A case of spatial exclusion is a simple generalization such as *There are fish in the lake*. Instead of pinpointing the location of the fish exactly, we tend to choose a more salient, larger region within which they occur, thereby facilitating both production and interpretation of the expression. The reason that the most natural size of the region[118] coincides with the boundaries of the lake is that the lake is in normal circumstances treated as an *individual entity*, though this depends on the chosen scale. We do not normally think of this as a generalization, but in fact there are obviously not fish at all points within the lake but only at certain spots, which are relevant exactly because they are the spots where the fish are. Like topics, such generalizations are not only bisubjective but circular and depend on the concept of relevance. Consider an example from classificatory exclusion, which is a more commonly treated subject than spatial exclusion: *If I were a bird, I'd fly there.* This would certainly not be an option if the speaker were an ostrich, but ostriches are implicitly excluded from consideration in this context by reason of being irrelevant (and they are irrelevant exactly because they do not fly).

Because generics are generalizations over perspectives, they neutralize the concerns about whose perspective the expression represents. In other words, a topic label taken on its own refers to all

[118] This region is also known as the *search domain* (Hawkins, 1984), a term that is rather unfortunate in an IR context.

relevant instances of the type and the response should hence be equally relevant for anyone interested in that topic. However, it is important to be aware that what instances are included in the topic as being relevant is a subjective choice and hence the interpretation of topic labels does indeed depend subjectively on the user. For instance, a user issuing the query *birds* with the intention of learning about flight techniques implicitly excludes ostriches from that concept for the purposes of that search, yet the system can hardly know this. Even generic topic labels are thus highly subjective, and their interpretation depends on context and intention.

5.4.2 User perspectives

Let us sum up some of the ideas presented here as they apply to search. Since the *intermediating user* (cf. section 2.4.5) is the one who formulates the actual query based on the requesting user's needs, it corresponds to the linguistic *observer* in the theory of reference suggested by Alfort (2009) and introduced briefly above, i.e. it represents the rational subject behind an utterance. The requesting user is the one who experiences the situation directly and thus corresponds to the linguistic *host*. The observer then adopts a point of view based in the host, which is basically a location, though often that of a person. Prototypically, the world is viewed from the observer's own point of view, but this is in no way requisite; the situation can be described as if seen from any point of view. Search, too, can be performed on the basis of a different point of view from one's own; for instance, a user may search for information that her friend needs, in which case she is the *observer* (in referential terms) and the *intermediating user* (in the suggested IR terms), while her friend is the *host* and *requesting user*. You can also search as either yourself or someone else at a different point in time or in a hypothetical situation. Imagine, for example, that you will be going to Valencia in three months' time and you decide to search for places to stay in advance. In this case, you (as a person in the present) are the observer and the intermediating user conveying the need by issuing a query, while the future (and thus hypothetical) *you* that is located in Valencia is the host and the requesting user who has (or is imagined to have) the need. This is of course not conveyed by the query, and so is a piece of information that needs to be explicitly provided by the user if there is a conflict of interests. In context-aware systems such as are currently in use in portable devices such conflicts may indeed arise, because the location on which you want the relevance of the information to be based differs from your current whereabouts. There may even be incorrectly resolved ambiguities, such as homophonous place names, where the one closest to your current location would be assumed to be meant, and ambiguous terms where you are currently interested in a different sense of the word than would normally be the case. For an example, imagine a software programmer who receives a call from a friend on vacation in the tropics. The friend is staring at a snake, and the programmer looks up *python* in order to see whether that is likely to have been the species in question. Here, a user profile would tend to assume that the programmer would be interested in a different sense of the word *python*. Naturally, this could be claimed merely to be a case of a short-term interest conflicting with the long-term ones. However, if looked at from a different angle, short-term interests may be considered changes in point of view, or in other words the host and requesting user is temporarily switched to a different person. Here, "person" should not be taken to mean a physical individual. Rather, the same person may have certain interests as a professional and others as a private person, and these may be considered different users residing in the same body. The same individual can thus represent different users at different points in time. In fact, this may concern the very conceptualization of the surrounding world, i.e. the ontology behind the word choice. For instance, in their professional capacity biologists may insist that bananas are berries, yet in their daily lives they may nonetheless call them fruits. In such cases, the word *fruit* would have a different meaning depending on the role that the person in question was taking on at the moment of search, i.e. there would be two different requesting users involved.

5.4.3 A bidirectional set of evaluation measures

A user searching for *dogs* would be likely to consider documents containing the following references to be far from equally relevant.

5.10	**Dogs** make wonderful pets.
5.11	He chased **the dog** away.
5.12	Photo of Christine holding **her dog**.
5.13	It is important to keep **your dog** properly entertained.

As can be gleaned from these examples, term reference makes a crucial difference; it is especially important whether the reference is to a specific dog or to dogs in general (or a generalized example of the class). Hence, reference and referentiality are of key importance for the relevance of a text, given a certain topic. If the user wishes to find information on *dogs* in general, then the tool should provide documents treating dogs in a generic way and ignore those concerning specific individuals. This requires an understanding of the intention of the author of the document as well as that of the user. It is not immediately easy to know when the word *dogs* in a query refers to the generic type and when any representatives of the class would answer the user's needs. It is somewhat less problematic to distinguish between generic and specific (and non-specific) uses in a document, because full text provides more linguistic clues. This is not to say that it is easy for a computer to decode natural language, however. The point is merely that the evidence on which to base such an analysis is far scantier in the case of queries than in full text. The material is in any case highly different in the two cases; for the author of the document, we have more linguistic material, but we cannot interact online with the author in order to resolve ambiguities. In the case of the user, we do have this opportunity to a certain degree, and we need it, because a query is very scanty evidence of a user's intentions indeed.

Sometimes, other words may imply greater relevance to the user searching for information on *dogs* even than some of the uses of the word *dog*, as in the following examples.

5.14	This site is dedicated to **golden retrievers**.
5.15	Always wanted to acquire **a golden retriever**?

After all, golden retrievers are dogs, and thus for some tasks these sites might be just as relevant as ones containing the form *dogs*, and far more than some of those containing *dog*. There is thus both generality and vagueness in documents as well as in queries. We may accordingly ask ourselves whether a document containing the word *illnesses* implies that the information concerns all illnesses and hence that the document should be retrieved whenever the query refers to an illness, or whether it is only relevant when general information on illnesses is required. In other words, given two documents of which one contains the word *illness* and the other the word *diabetes*, sometimes the query *illness* will make both documents relevant and sometimes only the former, while sometimes the query *diabetes* will make both documents relevant and sometimes only the latter.

Establishing the proper (intended) hierarchical scope of terms in documents and queries is an extremely challenging undertaking; one attempt has recently been made by *Yan et al.* (2011). Current IR system evaluation measures do not permit us to judge to what extent a system is able to return relevant and only relevant resources in this regard because they do not acknowledge the difficulties associated with matching two terms both of which can be interpreted in more than one way, such as in a generic or specific sense. The theory that allows us to understand such *bisubjective* matches has been developed above, and this section adds a set of bisubjective evaluation measures.

A truly subjective measure of search quality must take the interpretation of queries as well as of documents into account. For this task, any evaluation measure that one chooses to use must be applied to different test sets so that more than one measure is calculated. **Document interpretation** (**DI**) can be measured by keeping the query constant (as in an information extraction task where the question is clearly defined) and comparing the system's performance to manual annotations. A decision must be made whether a broad term such as *illnesses* is to be interpreted as including subordinate concepts or not. This might be called the **fixed interpretation** chosen for evaluation purposes. The task of the system is then to interpret the terms occurring in the document so that a search for *diabetes* will also return documents on *illnesses* if and only if that reference is judged to be relevant to all subordinate concepts rather than being information on the general concept only. This is traditional relevance evaluation, albeit in a sophisticated form. **Query interpretation** (**QI**) can be measured by keeping the target text interpretation constant. For the sake of this test, all terms in the documents are interpreted either as meaning exactly what they say (i.e. they are not expanded into subordinate concepts), or as encompassing all subordinate concepts as well. The system should then be able to decide whether the query *illness* is supposed to mean illness in general or whether it includes all subordinate concepts in the current context. Evaluation measures can now be applied according to one query interpretation. If the query is interpreted inclusively then all documents on illnesses should be retrieved. If it is interpreted exclusively (i.e. generally) then only the general information should be returned. This measure is not usually performed, and thus the ability of the system to adapt to different kinds of users with varying needs is not captured. Consequently, half of the problem of matching queries and documents is not even considered. To illustrate the methodology of query interpretation evaluation, QI precision and recall can be defined in the following way.

<u>QI Precision</u> is calculated as the number of queries in a search log which are correctly interpreted as having the chosen fixed interpretation (R) divided by the total number of queries deemed by the system to have the interpretation in question (s).

$$QIP = \frac{R}{s}$$

<u>QI Recall</u> is calculated as the number of queries in a search log which are correctly interpreted as having the chosen fixed interpretation (R) divided by the total number of queries in the search log that actually have this interpretation (r).

$$QIR = \frac{R}{r}$$

These formulas can of course be expanded into related alternative measures, such as the F-measure[119].

$$F = \frac{2 * QIP * QIR}{QIP + QIR}$$

Likewise, the generalized recall and precision measures can be stated in the following manner.

$$gQIP = \sum_{q \in R} r(q)/|R|$$

$$gQIR = \sum_{q \in R} r(q) / \sum_{q \in Q} r(q)$$

Where *q* is a query in the log *Q*, *R* is the amount of queries interpreted as having the fixed interpretation, and *r(q)* is the graded relevance value assigned to *q*.

This measure introduces a graded relevance value expressing the degree to which a query has the fixed interpretation. This interpretation may in fact not be very clear-cut, so it may be easier to decide which one of two interpretations is most like the one intended by the user, but not whether it actually matches exactly.

The fact is, of course, that probably no existing system is able to distinguish between intentions to find information on the general topic and intentions to include subtopics[120]. However, if we wish to improve search engines on this point, new evaluation measures are going to be needed.

If anything, this chapter has shown how much work we still have to do if we want to be able to have a machine interpret language, let alone disconnected query terms.

> Much good work applying the insights of the philosophy of language to information retrieval remains to be done. (Blair, 2003)

However, there is no reason to despair just because the morsel needs a lot of chewing. We now return to the task of producing a model of relevance in order to be reasonably clear about what users value in a resource before turning to actual query analysis in Part III.

[119] The literature abounds with papers on various evaluation measures, which have not been introduced here for reasons of space. For more information see Tamine-Lechani, Boughanem and Daoud (2010), Luxenburger (2008), and Kekäläinen and Järvelin (2002), among many others.

[120] It can be exceedingly hard to determine whether general information (e.g. on illnesses) also qualifies as particular information (e.g. on particular illnesses). Furthermore, having established that it does indeed concern individual illnesses, the question arises whether it is relevant in relation to *all* illnesses or merely prototypical ones.

6 Modelling relevance

> *Studies of relevance in information science have come a long way from simplistic assumptions and theological pronouncements. By now we know so much more about this complex and delightfully human notion. But there is so much more to learn, understand, and explore theoretically and observationally. Pragmatic improvements in IR, as it is constructed, depend in large part not on better and more sophisticated technology and networks, but on better understanding of relevance, and on incorporation of such understanding in IR processes.*
>
> (Saracevic, 1996)

We now return to the main purpose of this part of the thesis. This chapter is an attempt at opening the "black box" of relevance and satisfying a long-felt need by introducing an overarching model of the phenomenon. Saracevic (2007) summarized the important characteristics of relevance identified in the literature so far in the following way.

Saracevic (2007)

- **Relation**: Relevance arises when expressing a relation along certain properties, frequently in communicative exchanges that involve people as well as information or information objects.

- **Intention**: The relation in expression of relevance involves intention(s) – objectives, roles, expectations. Motivation is involved.

- **Context**: The intention in expression of relevance always comes from a context and is directed toward that context. Relevance cannot be considered without a context.

- **Internal context**: Relevance involves cognitive and affective states.
- **External context**: Relevance is directed toward a situation, tasks, problem-at-hand. Social and cultural components may be involved as well.

❖ **Inference**: Relevance involves assessment about a relation, and on that basis is created or derived.

❖ **Selection**: Inference may also involve a selection from competing sources geared toward maximization of results and/or minimization of effort in dealing with results.

❖ **Interaction**: Inference is accomplished as a dynamic, interacting process, in which an interpretation of other attributes may change, as context changes.

❖ **Measurement**: Relevance involves a graduated assessment of the effectiveness or degree of maximization of a given relation, such as assessment of some information sought, for an intention geared toward a context.

The Regulated Flux model suggested here takes all of these observations into account and elegantly assembles them into a coherent whole. It will be seen to be a highly complex matter requiring the recognition of the interaction of six dimensions[121], a hierarchy of three relevance spheres, six resource facets[122], and a set of three bisubjective gradation scales. This adds up to hundreds of different measures of true relevance in addition to the countless possible proxies for these. It is no wonder, therefore, that it is hard to find a suitable way of measuring the success of a ranking algorithm. As each of the individual measures is highly specific, few of them would ever be suggested as *the* measure of relevance (though various sets of them often are, depending on the use to which the notion is being put); it is only when they come together that they make sense as constituents of the general relevance of a resource.

The definitions of relevance types suggested in the literature often conflate certain criteria, sometimes for a good reason and sometimes less so. One exception to this tendency to conflation is topicality, which is often thought of in a very narrow way corresponding to what in the suggested model is the

[121] Possibly even eight (cf. Section 8.3.4).
[122] Again, this number may possibly be increased to eight in future.

cross-section of *focal relevance* with *topicality*, graded along the scale of relevance *strength*. This is just one cell in the huge matrix of relevance criteria. It is certainly an important one, but a much broader treatment is imperative if we are to acquire any kind of understanding of what is really going on in relevance.

The model presented in section 6.1 below consists of a set of interdependent dimensions which together result in the judgments that can be observed on the surface, i.e. it is *stratified* in the sense of Saracevic (2007), though the actual strata differ. This accounts for the complexity, and also the dynamicity, of relevance when considered from outside. Dynamicity across the search process is mirrored by the different stages to which the various dimensions are assigned; at each stage, another type of relevance is given a final assessment, until all criteria have been verified (or at least estimated). This temporal association of the dimensions leads to hierarchical relationships and different strengths of dimensions; not all criteria carry equal weight for a given user.

Saracevic notes that, in his model, "the strata are not necessarily imbedded within each other, nor do they form a hierarchy. The relations between strata are much more complex and could be in flux" (Saracevic, 2007, p. 1926). The Regulated Flux model of relevance explicitly recognizes this by providing principles which regulate when flux is permitted to occur, i.e. at which stages users are allowed to update their relevance assessment values concerning particular relevance dimensions. The relevance dimensions comprise what shall be termed the *hierarchy of relevance dimensions*, which can be shown to correlate systematically to the *hierarchy of relevant contexts* identified above. In fact, this correlation of the two hierarchies may be considered the very essence of relevance; it consists in a relationship between the spheres of interest surrounding a rational being in a situation in the world and the corresponding features of a resource. This is discussed in depth in section 6.2.

Section 6.3 looks closer at relevance gradation along various scales and deals with the bisubjectivity of relevance assessment. In each dimension, relevance can be graded along (at least) three different bisubjective scales. This accounts for the many ways in which relevance can be said to be a matter of degrees. From the human point of view (i.e., to the exclusion of pure and uninformative matching of strings), there is not a single type of relevance which is not gradable in some way. There can be many reasons for preferring a certain resource over others. For example, the length of the document, the availability or price of it, the style in which it is written, and even the political, religious or theoretical framework in which it is written may be decisive for the user's interest in it[123].

[123] It is important that ideological filtering is not carried out without the user's knowledge, because that may affect the user's views in inappropriate ways (Teevan, Alvarado, Ackerman, & Karger, 2004).

The model is brought to the test in section 6.4 by applying it to assessment criteria identified in behaviour studies. The criteria that users report from actual relevance assessment are difficult to interpret, because they often correspond to complexes of several relevance dimensions, and frequently refer to second-order causes, i.e. the causes of characteristics that in turn trigger relevance. However, if one recognizes these difficulties, the empirically reported criteria can all be assigned to the six dimensions. As the model is able to explain why those criteria are chosen by users and why they make resources relevant, it is concluded to be a successful model of the observed phenomena.

Apart from the model, a definition of relevance for the purposes of IR is attempted in that same section. As far as possible, it has been endeavoured to make the model and the definition suitably general to be useful in discourse studies as well as IR. The model needed for IR draws on the results arrived at in discourse studies, but it might in turn be applied to communication in general in order to enrich the understanding of relevance in a broader sense. The two research fields have similarities as well as differences, and both of these circumstances call for cross-disciplinary discussions; the similarities mean that researchers from the two camps can draw on each other's results, and the differences that they can learn from one another's different data (section 6.5). Because the default functions and purposes of an utterance are so different from those of a query, the two research fields have developed different foci and have overlooked different types of relevance. Here, they are brought together in one model, the Regulated Flux model.

6.1 Regulated Flux – A unified model of relevance

It would be entirely uncontroversial to suggest that in a topic search relevant information is that which is capable of changing the user's understanding of a topic. It follows from the various discussions above that this reduces to a requirement of informativity and topicality on the part of the resource. This reduction of our definition of relevance in topical search into constituent phenomena shows us that relevance is composed of several building blocks. These are the **dimensions** which may co-occur or appear separately. The dimensions are closely related to Saracevic's *manifestations*, but that term is dispreferred here because it may suggest that only one manifestation is present at any one time when in fact all dimensions are at play simultaneously.

6.1.1 Relevance dimensions

It may be a good idea at this point to briefly recapitulate the essential topic spheres based on Schütz's ideas.

- **Focal relevance**: Something is selected by the individual as being of current interest.

- **Local relevance**: Something is perceived as being problematic and is distinguished from the horizon as a theme on which to concentrate.

- **General relevance**: Something is judged as being relevant on the background of the horizon, the background knowledge possessed by the individual and the past experiences.

It is important to realize that not only focal topics but all of Schütz's spheres of relevant topics represent topicality. The traditional portrayals of topicality, utility and interest may easily lead one into thinking that only focal topics are truly topical (i.e. on the topic of the query), when in fact they all represent topics that are relevant to the person in question in different ways. Also, one might think that local relevance represents utility pure and simple when in fact focally relevant documents may obviously likewise be useful. If the user needs information on a certain topic without reference to more concrete application needs, then information on that topic is indeed useful apart from being topical in the sense of focal relevance. Similarly, it goes without saying that focally and locally relevant topics may be just as interesting to the user as the generally relevant. Thus, the topical spheres are independent of the distinction between topicality, utility and general interest; they are simply topical spheres of attention.

Relevance is not just about topicality, utility and interest, however. As should be clear by now, we shall in fact suggest a model in which there are six dimensions as follows, corresponding to the six facets of a resource[124] and completely independent of the Schützian spheres, which will, however, enter the scene again as soon as we start focusing on the various dimensions, which may all concern either of the spheres.

[124] The reader will recognize the numbering with Roman numerals from the discussion of facets in Part I.

I. **Intentionality**: Was the resource intended to be read by the user or someone like the user?

II. **Practicality**: Can the resource be accessed by the user in a readable format and read in a language known by the user?

III. **Informativity**: Is the information content able to change the user's understanding and provide cognitive gains?

IV. **Topicality**: Is the content on a topic that is currently considered relevant by the user according to the HRC?

V. **Utility**: Is the resource helpful in performing a task (including the simple collection of information)?

VI. **Desirability**: Does the user consider the resource satisfactory and appealing?

These six dimensions can be translated into a set of circumstances that the user is in at the time of search: The user belongs to a certain user group (*intentionality*), has certain possibilities (*practicality*), a certain amount of previous knowledge and a given ability to understand (*informativity*), an idea of what topics are of current interest and how these are delimited (*topicality*), a task to be performed (*utility*), and certain motivations, preferences and interests (*desirability*). On the other hand, resources also possess certain features that pre-code them for relevance: They were intended for certain readers (*intentionality*), are provided in certain formats (*practicality*), contain a given quantity of information presented with a certain amount of clarity (*informativity*), concern a set of topics (*topicality*), are useful in certain types of situation (*utility*), and have certain qualities (*desirability*). The IR system should compare these two sets of initial states, and also monitor any changes in them (in particular that representing the user) along the way.

Though simplistic in the form in which it is presented above, the list is intended to be exhaustive, but it is of course possible that additional dimensions may have escaped notice. However, it should be noted that as each dimension corresponds to one of the six facets of a resource, the Faceted Stimulus model should most likely be amended correspondingly if any new dimensions were added. This is in fact a point where the model – or a future amendment of it – may be tested; if a dimension makes sense as a kind of relevance but not as a facet, or *vice versa*, then there is likely to be something wrong with the model.

The dimensions do not compare to notions such as logical or objective and situational or subjective relevance; they are all subjective. In fact, all dimensions are *bisubjective*, i.e. they involve two points of subjectivity: The user requires a certain level of relevance, and the resource satisfies a criterion of relevance to a certain degree. All dimensions also incorporate a trace of objectivity, especially in a negative sense, i.e. some resources are clearly irrelevant according to a given criterion. This is not to say that some types of relevance are objective, but rather that given certain subjectively defined conditions some relevance dimensions are blocked. For instance, given an assessment that a resource is not on a useful topic, its informativity value can be set to zero, because informativity on an irrelevant topic does not contribute to relevance. More will be said about this in section 6.2.1. The following is an overview of the six dimensions of relevance as suggested here.

● **I – Intentionality**

Some resources we have to read. Maybe we have been told so by our superiors, or we just feel that we need to have read those resources. In this case the intentions are not necessarily those of the author, but the informer in the guise of someone who at the very least told us to get hold of the resource and read it, thereby making it relevant. If someone intended the user to read a certain document it is relevant to him or her by definition. Other dimensions, such as topicality, may of course rank higher in the eyes of a particular user, but the stimulus document cannot be said to be completely irrelevant.

Intentionality frequently appears even before a resource is accessed. A resource is high in intentionality if someone intended to inform the user or someone like the user in a situation similar to that of the user. It is low if it was not specifically intended to inform someone like the user or to provide information applicable in that kind of situation. Sometimes, the feeling that the user needs to read a certain resource may appear during the search process, as when a participant in a user study explained that *I'm nervous about this article because I was really hoping nothing had been done, so I really need to look at it* (Maglaughlin & Sonnenwald, 2002).

Clearly, it is the intentional facet of a resource that is judged with this dimension. The Schützian spheres may be applied to intentionality as well as topicality. In this case, **focal intentionality** is a matter of whether the text was explicitly intended to be read by the user in question, **local intentionality** is present if the resource was intended to be read by someone in a situation similar to that of the user, and **general intentionality** if it was intended to be read by anyone in a more general group that includes the user.

● II – Practicality

The first thing a user must determine after intentionality assessment is whether the resource is at all available, and at what cost. Of course, in practice this step may be gone through after it has been decided whether it is on the right topic since topics can to some extent be identified from titles and snippets, but in principle at least acquiring the resource is the first step in the process. A resource which is high in practicality is easily obtainable and readable without too much cost in effort, time or money. This is independent of actually understanding or accepting the contents; efforts spent on this belong with informativity. If a document is not obtainable, or is obtainable in an unreadable format or language only, it is irrelevant as concerns practicality, though it may still be (potentially) relevant for other reasons, i.e. along other dimensions. Practicality may seem a highly prosaic dimension, but it is nevertheless among the criteria that people mention in relevance assessment. Some authors (Hertzum, Andersen, Andersen, & Hansen, 2002; Borlund, 2003; Xu & Chen, 2006) are of the opinion that practicality criteria "are not directly concerned with the user's information need, and consequently, do not fit into the definition of relevance as a relation between corpus of documents and different (identifiable) aspects of an information need" (Borlund, 2003, p. 918), and hence that other criteria should be the attention of future research. However, I fail to see how the accessibility of information can be considered unrelated to the user's information need. It is simply the dimension that takes care of the substantial facet of a resource, which is just as important as the others, even if it is somewhat less exciting. It is suggested here that a very broad notion of relevance be adopted in order that the interaction of all types of relevance assessment criteria may be explained and understood. Authors of studies may then freely exclude some relevance dimensions from consideration in specific cases if they so wish. It is of course debatable whether that broad phenomenon should be called *relevance* or something else, as argued by Janes (1994) for instance, but to me it seems only natural to collect it all under that name, despite the fact that "that word too carries baggage in people's minds" (Janes, 1994, p. 168)[125]. Otherwise, the notion would have to be restricted to one dimension only, and could not, for instance encompass both topicality *and* utility, since that would be an arbitrary choice of dimensions taken from the whole that we

[125] Janes could not come up with a suitable name for the overarching phenomenon and ended up simply calling it *the Big Black Question Mark* and leaving the choice to future researchers.

shall therefore continue to call relevance. This is in accordance with the use of the term by other authors who have tried to reconstruct the phenomenon in its entirety, such as Schamber *et al.* (1990) and Saracevic (2007).

Once again, the Schützian spheres interact with the dimension of practicality: **Focal practicality** deals with whether the user is allowed access to the resource in question, **local practicality** whether it is readable by someone like the user (i.e., in an appropriate script, etc.), and **general practicality** concerns the availability of the resource at a certain cost.

● III – Informativity

The next step in the process of accessing a resource is extracting the information it holds. This can be easy or demanding, depending on its contents and the user's background. Previous knowledge on the topic or experience with the phenomena described may facilitate understanding. Informativity is high if the information is capable of changing the user's understanding of a topic without too much cost in processing efforts, and the information is accepted.

A potential to inform does not in itself cause relevance; relevance is subjective and depends entirely on the user. If the user does not trust the source (e.g. an author or a journal) or doubts the veracity of the information, he or she may not accept it. This will block the informative potential of the resource and make it irrelevant as concerns informativity. The resource may, however, still be relevant according to other criteria. Sometimes, the fact that the user does not agree with the information, or even does not like it or dislikes the author may cause the same effect; the user will not allow the information to change his or her understanding of the topic, and hence informativity is blocked. A bad presentation or unclear linguistic expressions may also lower informativity. Furthermore, informativity depends on the sheer amount of information provided in the resource. If the user already knows the information in question, this will also lower informativity in spite of the informative potential of the resource.

Informativity represents the resource's effect on the user's understanding by combining information on what the user already knows and how well the resource achieves its mission of informing the reader. This dimension is highly subjective in that it both refers to the state of the user's mind and to the interpretation that the user is making of the resource information.

Informativity is the dimension that evaluates the formal facet of the resource. **Focal informativity** deals with the user's ability and willingness to extract the needed information from the resource, **local informativity** is about the ease with which information can be extracted by anyone like the user, and **general informativity** concerns the extent to which the information is at all present.

● IV – Topicality

Topicality is the dimension that takes care of the representational facet of the resource which deals with the semantic contents. A resource is topical if the information is on a relevant topic. It is generally assumed that this relevant topic is what the query dictates. However, with Schütz' types of relevant contexts we can now define three topical spheres which may all be said to be relevant to the user in different ways and with varying urgency: Focal, local and general topics. The query supplies a **focal topic** (or at least, is a proxy for one), while **local topics** are those that will actually solve a problem or help with a task. **General topics** are just those which the user will find interesting but where information is less urgently needed. Local topics might in principle be accessible through some kind of situation description, if it could be determined what kind of information would be useful in that kind of situation, while general interests are stored in long-term user profiles. Short-term user profiles correspond to local and focal topics. In this interpretation, then, topicality does not necessarily imply that the resource in question is on the topic described in the query. Rather, it means that it is on *a* topic related to one of the three interest spheres. This is a substantially different interpretation than the one usually adopted, but the present author believes this is a necessary development given Schütz' findings.

A resource is high in topicality if it is clearly on a contextually relevant topic, and low if it is more remotely connected to it. Often, remote topicality will require much processing effort to identify. In principle, if the document is not on any relevant topic then it is irrelevant as regards topicality. However, it is hardly possible in practice to determine with certainty that a resource is topically completely irrelevant (Cooper, 1971). A complete lack of topicality would block utility; by definition a resource cannot be useful if it does not contribute anything of value within the spheres of relevant context. The user defines subjectively what the topic in question comprises, and also interprets what the resource information seems to be about. Note finally that topics may extend in the spatial, temporal and classificatory dimensions, i.e. a resource may fail to be relevant because it is too old (or new for that matter), valid for the wrong region, or about the wrong kinds of phenomena.

Only when the information has been read and accepted can it be determined whether it is on a topic of current interest. Of course, as with practicality, in practice this may happen the other way round, because the user may have decided from a snippet or even the title of the document that it is indeed on the topic, before even attempting to read the actual text.

Sometimes, topicality is defined by collective beliefs such as scientific "truths", which may or may not be uniform across all scientists within the field. Thus, as pointed out by Hjørland (2010):

> [If] you believe schizophrenia is caused by double-bind communication in childhood, then family studies – and the information they can provide – become relevant. If, on the other hand, you believe that schizophrenia is caused by chromosome disorders, then genetic studies become relevant. (Hjørland, 2010, p. 229)

Topicality is thus highly subjective, as discussed at length in section 4.5.

● V – Utility

The next step in the relevance judgment process is *utility*. This assessment is performed after the information has been incorporated and the topic ascertained. At this stage, the information is interpreted in the context of the task that prompted the search. If the information is not found to be applicable to the task, it is not useful in the sense of utility. This may be caused by a changing understanding of the task and of what is needed during the process of incorporating the knowledge. Utility thus realizes the dynamicity of relevance.

Originally, there was a strong focus on topicality in IR studies. Later, it was realized that resources may be relevant in spite of not being on the topic dictated by the query. Hence, utility-like relevance types were conceived. These consequently focus on usefulness in the task situation. However, utility can in fact be said to be present not only in local but also in focal relevance (information is useful if the task is simply to understand a focal topic), as well as in general relevance (information is useful in entertaining or increasing the user's general understanding). As with the other dimensions, there are thus three Schützian spheres of utility. **Focal utility** is about the resource's ability to fulfil and satisfy the need, **local utility** concerns whether it is helpful in the present situation, and **general utility** implies general usefulness irrespective of the current state of affairs. Clearly, like topicality, *utility* in this interpretation is much broader than originally intended. A resource is high in utility if it is easily applicable to the task and the user gains something useful from reading the information. It is low if the user finds that it did not satisfy his or her needs. Like the other dimensions, utility is doubly subjective in that it involves the user's interpretation of what is needed in the present situation as well as of what the resource information might be useful for.

The utility dimension corresponds to the interpersonal facet of a resource, and hence it realizes the communicative aspect of a document. While the contents may be topical, what the author succeeds in communicating and the ways in which the author is trying to help the reader belong with utility. Another obvious and critical criterion belonging with utility is whether the user already has access to the resource and hence has no need for another copy. This is another example of a judgment that precedes its logical place in the process, since it is likely to frequently take place before the resource is accessed.

⬬ VI – Desirability

A resource is desirable if it has aesthetic appeal or inspires interest; in other words, if the user is inclined to access, read and apply the information[126]. Again, this is usually decided at a much earlier stage. It does, however, represent the ultimate decision on the part of the user with respect to the relevance assessment.

The model proposed here assumes that in the end, what the user wants is what is most relevant. This is not as self-evident as it might seem. Some authors are of the opinion that users are not qualified to judge relevance unless they happen to be experts in the particular topic they are researching (Hjørland, 2010). In other words, an IR system may suggest information that is truly relevant even though the user is unable to detect its relevance[127]. However, users can hardly be helped by reading such documents. The crucial point is that relevance according to the suggested model is only ever fixed when the last benefit has been had. Thus, by definition, if the user ends up by considering the resource relevant, it will be relevant. If, on the other hand, he or she never realizes its potential, it will remain irrelevant. Such a model succeeds in explaining how relevance works in practice. It does not, however, help IR systems provide information which "is truly relevant without the user realizing it". For such tasks, separate methodologies are needed. The claim made here is not that this would

[126] I.e., access the substantial, read the formal, interpret the representational and apply the interpersonal resource.

[127] As mentioned by Hjørland (2010, p. 226), this may for instance be a very good idea with medical information, because "the user normally does not have qualifications to evaluate his or her information need". However, that seems to be a special case; when going to the doctor patients explicitly expect to receive suggestions beyond what they might propose themselves. With information search in general, this approach may not meet with quite as much enthusiasm.

not be recommendable but merely that providing users with relevant material in excess of what they are able to appreciate as such is not part of a relevance model but a question of teaching and learning. It is certainly an important and exciting task for others to explore.

In its general sense, desirability corresponds to what is often called *satisfaction*. The dimension deals with the effective facet of a resource, i.e. whether users are pleased with the way the other facets affect them. Thus far, only the emotional aspect of desirability has been illustrated, but there are more tangible sides to it as well. If the actual contents do not fit what the user had in mind, in spite of being topical, it may be considered irrelevant. This happens if the user was interested in a description from a different point of view, or an image of a different colour, etc. (Choi & Rasmussen, 2002). Apart from the emotional reactions this may cause, and the possible resultant influence on informativity, it can also simply mean that the user rationally considers the resource unsatisfactory and keeps looking for one with the desired qualities. The reader may be surprised that such rejection of resources that do not fit what the user had anticipated should take place after utility evaluation, but in fact this makes perfect sense; the resources in question are perfectly useful. It just so happens that the user prefers something different if at all possible. He or she may even revert to the rejected resources if no better alternative is found, cf. the responses *Once I saw the previous photo judged as more useful, this photo turns to be less useful since it had a little bit different perspective* and *This photo is less useful. But it is important because of the difficulty of finding other relevant pictures* reported by Choi and Rasmussen (2002). If, on the other hand, a resource is entirely useless with the wrong characteristics, then it is also a matter of lacking topicality and hence is likely to be rejected earlier in the assessment process.

Focal desirability implies that the resource is felt to be relevant by the user, **local desirability** that it is suitable as an answer to the request, and **general desirability** that the user finds it appealing. Already O'Connor (1968) recognized this trichotomy of meanings assigned to the notion of *satisfying an information need* and saw it as a problem to be solved. We shall not consider it in the least problematic; satisfaction may simply relate to any of the three spheres. Desirability is perhaps the most obviously subjective dimension. It may, however, not be so clear that it is also doubly subjective like the other dimensions, though it does in fact involve both the identification of what appeals to the user and the judgment of whether a given resource satisfies this criterion of satisfaction, aesthetics and interest.

It is hoped that the above will have illustrated sufficiently clearly that relevance is a complex of co-occurring dimensions, all of which may contribute to the relevance of a resource, and that the

Schützian spheres make for an even more multifaceted phenomenon than usually believed, even by those who maintain that it is complex (fig. 6.1).

	Focal	**Local**	**General**
Intentionality	Intended for the user	Intended for someone in a similar situation	Intended for anyone like the user
Practicality	User is allowed access	Readable by someone like the user	Resource is available
Informativity	User is able and willing to extract information	Information may be extracted (easily) by someone like the user	Information is present
Topicality	On the topic asked for	Information on the topic is useful in the circumstances	Information is interesting
Utility	Fulfils the need	Is helpful in the circumstances	Is generally useful
Desirability	Is considered relevant by the user	Is suitable in the circumstances	Is generally appealing

Fig. 6.1. An overview of the six dimensions and the Schützian spheres with which they interact.

6.1.2 Affective characteristics of relevance

The terms *affective* and *motivational relevance*, suggested by Saracevic (1996), are very vague and abstract, reflecting the difficulty that researchers have had in identifying the exact effects of this aspect of relevance[128]. Borlund (2003, p. 915) dismisses both entirely, pointing out that "the 'drive' to want information is not an independent, specific type of relevance, but an inherent characteristic of relevance behavior in general", and Cosijn and Ingwersen (2000) identify motivation as an attribute of relevance in general and affective relevance as a manifestation of success or satisfaction. Even Saracevic acknowledges in his 2007 definition of affective relevance that "it can be argued that affective relevance underlies other relevance manifestations, particularly situational relevance".

[128] See also Lopatovska and Mokros (2008) on affectiveness in information retrieval.

Thus, affection and motivation do not represent a separate dimension but are feelings activated during search. It is interesting that, taken generally, affective aspects of relevance cover a large proportion of the criteria that users apply, as shown by statistics from the user study performed by Barry (1994).

> Affection refers to emotional responses to any aspect of the document. A very interesting phenomenon is that 15.8% of the relevance statements by users in Barry's study fall within this group (compared to 35.1% on the contents of the documents, and 21.6% on the user's background and experience). Users therefore regard this type of relevance judgment as an important criterion overlapping other criteria in the selection of relevant information objects. (Cosijn & Ingwersen, 2000, p. 546)

The desirability dimension introduced here does not attempt to capture all kinds of affection and motivation, but is centred on aesthetic appeal and the inspiration of interest; satisfaction can be derived from all relevance dimensions, but some resources are more attractive than others, even to the point of being preferred despite being irrelevant according to most other criteria. However, the ultimate satisfaction of the need and the final accept of the resource as being relevant are features that do belong in the desirability dimension.

Note finally that the essential difference between utility and desirability is that the latter concerns the actual user personally and individually whereas a resource would be useful to *anyone* in the same situation and needing the resource for the same purpose.

6.1.3 Usefulness and pertinence

Corresponding to the different ways in which relevance deviates from the simplistic system model[129], several different kinds of subjective relevance have been posited in the literature. The most important types are *pertinence*, which is said to hold between a piece of information and an information need, and *situational relevance*, which holds between the information and the work task situation as perceived by the user. While pertinence requires an understanding of the user's needs that in itself exceeds what is possible in a traditional IR system, situational relevance necessitates even more contextual information than that. Situational relevance is often spoken of in terms of *utility* and *usefulness* in order to stress that the user decides what is relevant. Borlund (2003) suggests that it may not be possible to distinguish between pertinence and situational relevance in practice. However, this seems not to be true. As situational relevance refers to the way the work task situation is perceived by the user, the information need on which it is based is bound to be merely what the user *believes* that

[129] If indeed such a model ever existed among others than its very contestants (Hjørland, 2010).

he or she needs. Pertinence means relevance according to this perceived need, whereas situational relevance may go further and – given the right kind of contextual information – attempt to fulfil the *real* need (at least according to the task as perceived by the user). This will potentially result in entirely distinct responses. Needs appear in many guises, as explained at length in chapter 2, and deciding which need to attempt to satisfy is an important choice on the part of the system designer.

The difference between pertinence and situational relevance appears not to be one of relevance dimensions, but between topics in the HRC. Pertinence implies that the information is on a focal topic and also at least somewhat relevant according to all other dimensions, while for a resource to be situationally relevant means that it is on a local topic as well as having some relevance according to all other dimensions. To this could be added a third corresponding type which might reasonably be called **interest**, i.e. the relation holding between a user and information on a general topic (according to the HRC) which is also relevant according to all other dimensions. These are just three constellations of relevance parameters among many others, and recognizing the entire system of dimensions and spheres equips us much better for analysing relevance assessments.

6.2 Regulating the dynamicity of relevance

One might reasonably ask whether any of the dimensions of relevance are in general of greater importance than any other and thus should be assigned more weight. The answer must be that this appears indeed to be so; according to the HRC, what is of current interest should be given priority, followed by what is problematic, and only subsequently that which is generally informative or interesting. In other words, prototypical (semantic) topicality outranks prototypical (task) utility, followed by general desirability. How the remaining 15 types fit in is less obvious, and in any case the relative weighting is likely to vary between situations.

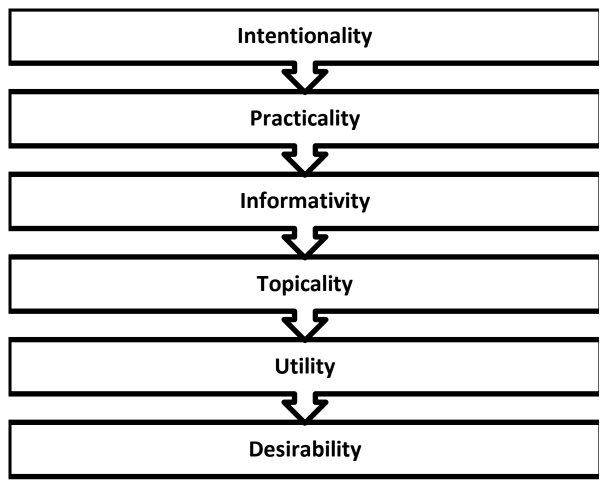


Fig. 6.2. The hierarchy of relevance dimensions (HRD). A resource high in a given dimension cannot be said to be relevant unless it is also relatively high in all dimensions below it.

6.2.1 The hierarchy of relevance dimensions

As noted in the description of the various dimensions above, there is another sort of hierarchy between dimensions in that some of them depend on others (fig. 6.2). This **hierarchy of relevance dimensions** (**HRD**) is related to the actual search process, and thus the model elegantly incorporates the dynamicity of relevance.

- **I – Intentionality**

Firstly, even though intentionality ensures a certain degree of relevance, a resource can hardly be said to be relevant in a more general sense (i.e. according to the ultimate relevance judgment) if it is not also practical, informative, topical, useful and desirable. For one thing, if it cannot be accessed, understood and applied to the task (as defined by a superior or other informer, including in some cases the user), or if the user simply refuses to read it, locating the resource is hardly useful.

As pointed out in chapter 4, it is important in this connection to realize that the relevance of suggesting a resource to a user is not necessarily the same as the relevance of the document as such, since any one dimension may potentially cause the reference to be irrelevant to the user in spite of the document being highly relevant as regards other dimensions. The task of an IR system is to calculate the relevance of suggesting a resource. Consequently, all dimensions must be taken into account in all cases.

- **II – Practicality**

Secondly, it is reasonable to suppose that a resource is not ultimately relevant just because it is high in practicality if it is not also high in informativity, topicality, usefulness and desirability; even if there is only a single document in a collection that is free, a user will not find that resource relevant if it is not in any way useful. Practicality is thus not enough, but it contributes to the overall relevance of a resource.

Another way of putting this is to say that what needs to be practical is the access to informative information, so the input to the practicality algorithm is the output information from the informativity test. Thus, if there is no informativity at all, for instance because it is written in a language or script that the user does not understand (or because there is no useful information on the topic to be informative, see below), there is simply no input to test for practicality, and thus this dimension outputs a null. This result in turn blocks all succeeding tests from being rerun, because if there is no practicality, there can be no informativity, topicality, utility and desirability. This also happens if practicality is zero for other reasons, such as the inability of the system to handle the document format.

● III – Informativity

Thirdly, an informative resource is hardly relevant if it is not also topical, useful and desirable. Even if there is only one resource in a collection that a user can understand, it will not be relevant if it is not in any way useful. A complete lack of informativity blocks topicality; if a resource does not provide information[130], it cannot contribute to any topic. Harter (1992) suggests that the notions of *relevance* and *information-as-process* are one and the same thing, and that consequently only one of the terms will be needed in information science. However, something may well be informative without being relevant to the question at hand, and it may even be relevant without being particularly informative. Informativity involves the quantity of information that the user is able to extract from a resource, with reference to its topic but not (directly) to its utility.

The model developed here explains these phenomena in the following way. The input to the informativity algorithm is the output information from the topicality test. Thus, if there is no topicality, there can be no informativity (a situation with no topicality whatsoever is of course not likely to occur very often), and so informativity will be null. This in turn blocks all succeeding tests from being rerun, because with no informativity no topical information can be extracted. This is also the case if informativity is zero for other reasons, such as the resource being written in a language or script that the user does not understand, or because of blocks further down the chain, such as a lack of access to the resource in question.

● IV – Topicality

Fourthly, it is reasonable to suppose that a topical resource is not relevant if it is not also useful in the sense that it is actually applicable to the task at hand. Topicality thus depends on utility and desirability. The reason that topicality is dependent on utility is that the topic is part of the use situation for which the information or other resource is needed; the situation defines the topic (note that we are talking of relevant topics in the actual resources and not the topics that users decide to search for, which of course have a much more indirect relation to the use situation).

The Regulated Flux model states that if there is no utility because the resource is not applicable to the task, topicality is set to null by definition, because if the resource is not on a

[130] Or what corresponds to information in non-textual resources.

useful topic then it is not topical. This in turn blocks the utility and desirability tests from being rerun, because with no topicality there is nothing to apply to the situation at hand or of which to judge the desirability. This also happens if topicality is zero for other reasons, such as a block further down the chain. For instance, if a text cannot be understood, the information cannot possibly be applied.

● **V – Utility**

If a resource is useful, it is usually relevant in spite of other dimensions being low. However, if the user chooses not to access a resource or read the contents because they are unappealing or uninteresting, the resource will end up not being judged relevant in spite of having a utility potential. If there is only one useful resource in a collection, but it is so unappealing that the user shies away from it, it is by definition not relevant since the user did not in fact want it; user satisfaction is the ultimate criterion of relevance, and satisfaction hinges on desirability.

According to the model, if there is no desirability (i.e. the user rejects the resource), then there is no utility either; it is set to null. This in turn blocks the desirability test from being rerun, and the resource is rejected as being irrelevant. This also happens if utility is zero for other reasons, such as the information being inapplicable to the task at hand, or a block caused by lack of topicality, informativity, practicality or intentionality.

● **VI – Desirability**

Desirability does not depend on any other dimension; the user may simply refuse to read a given resource, or may choose to access one for no reason at all except that it seems interesting (so-called *serendipity IR*). However, the desirability test may be blocked by a lack of utility. This is why desirability ends up being the ultimate relevance parameter; a resource can only pass the test if it has passed all others to some extent – though it may be very low in some dimensions as long as it is not zero (or null).

Clearly, the model allows for intricate interactions between the dimensions during the course of a search session; the cybernetic circle of relevance assessment runs continually, testing the stimulus resource for relevance against a model which is constantly updated until the final decision has been made which deems the resource to be either relevant or irrelevant. Mizzaro (1998) includes time as a separate dimension of relevance, but time does not in itself contribute to the relevance of a resource. Time is simply an external dimension along which relevance may change for independent reasons. The Regulated Flux model incorporates time in an elegant fashion by allowing tests to be rerun to a certain extent but not indefinitely.

6.2.2 The search process

The above walk-through is algorithmic and illustrates the rules of running relevance evaluation tests in the various dimensions. However, this is not the same thing as a description of actual user behaviour. The search process starts with an intention, goes on to the selection of a resource via letting it affect the mind and determining the value and usefulness of it to the final choice of whether to consider it desirable or not. Users being humans, some of these decisions routinely happen at different points in the process from what one might expect from a purely mechanical point of view. For instance, a preliminary estimate of topicality often precedes any thoughts of whether the resource is even accessible, because focal topicality is the centre of human attention, as pointed out by Schütz.

- **VI – Desirability**

  A resource immediately triggers an amount of desirability, even from the moment a snippet is shown in the results list. It is indeed possible that desirability is present even before the search is performed in the guise of motivation to find a suitable resource. If the user chooses to go on with the exploration of a resource, his or her opinion of it is then liable to change all the way through the process while reading the document and applying the information to the task. Consequently, the final judgment can only be performed at the last moment when no more stimuli are collected from the resource.

- **V – Utility**

  Whereas initial desirability assessment is unlikely to be conscious in most cases, initial utility evaluation seems to be the very mechanism that determines whether a user chooses to continue exploring. Thus, a guess is made (again as early as the moment when a snippet is first shown) as to whether the resource is likely to be useful. This opinion then potentially fluctuates throughout the process until a final judgment is made at the proper stage. After this point, utility assessment is fixed.

- **IV – Topicality**

  A guess at the probable topicality of the resource is likely to be made at an early point. However, this may well change during the reading process until the proper stage is reached where the final topicality assessment is made. After this point, topicality assessment is fixed.

- **III – Informativity**

  It is probable that a guess at the potential informativity of a resource is made when practicality is decided. If it is written in a language not understood by the user, then it is not likely to be very informative. However, the informativity evaluation naturally continues during reading, but not after the resource has stopped providing cognitive gains.

● **II – Practicality**

Practicality is fixed the moment the user attempts to access a resource, or even when the results list is presented, if it shows which resources are available to the user, and in which formats and languages.

● **I – Intentionality**

Intentionality is evaluated before the search is even performed. It cannot be changed by search results, though external events may affect all dimensions by altering the premises. The user's preferences may fluctuate, the task may change, a new topic may be chosen, the user's accept of a resource may be updated, the accessibility premises may change, and the intentions that the user should read a certain resource may be withdrawn. All such events are basically irretrievable for an IR system. On the other hand, the dynamic changes inherent in the process as sketched in this section cannot be ignored; they are invariably present, from the start of the search to the point when each is fixed.

6.2.3 The essence of relevance

Topicality seems to be a very strong criterion with focal relevance. It may be that if focal relevance fails and it is attempted to satisfy local relevance instead, utility takes over as the principal criterion. Similarly, if local relevance fails and the search reverts to mere general relevance, desirability may become particularly significant. Consider the example from before: If a user has searched for *train tickets Lisbon*, all documents bearing on this topic might be considered relevant, including one that informs the reader that there are no trains to Lisbon from his or her current location. If no document is found (or if the designers of the tool decided to focus more on utility, or simply to introduce more diversity in the results), the search may fall back on local relevance, i.e. solving the user's problem. Now plane tickets become highly relevant, but also other considerations related to utility, such as the applicability of the information to the task at hand. The focal topics as expressed in the query are demoted to mere guidelines to utility. If there are still no (or too few) results, the tool may potentially fall back on general relevance and suggest interesting documents related to the topics TRAIN and LISBON, and possibly to the need signalled by the query, including ones mentioning that it is possible to take a boat to Lisbon. This time, desirability criteria will be important; since at this point the tool is more or less "content" to entertain the user, or at least to retrieve material of interest to him or her, the appeal and entertainment value of documents will be of greater value.

This suggestion hints at a relationship between the HRC and the HRD. Desirability represents knowledge that *is liked* (general relevance), utility knowledge that *is needed* (local relevance), and topicality knowledge that *is wanted* or *asked for* (focal relevance). Informativity then corresponds to the knowledge that it is possible for the user to acquire from the resource, i.e. knowledge that *can be had*, practicality represents knowledge that *exists* or *is available*, and intentionality knowledge that *should be had*, according to someone's intentions. It is possible to extend the HRC to incorporate all these levels accordingly as in fig. 6.3.

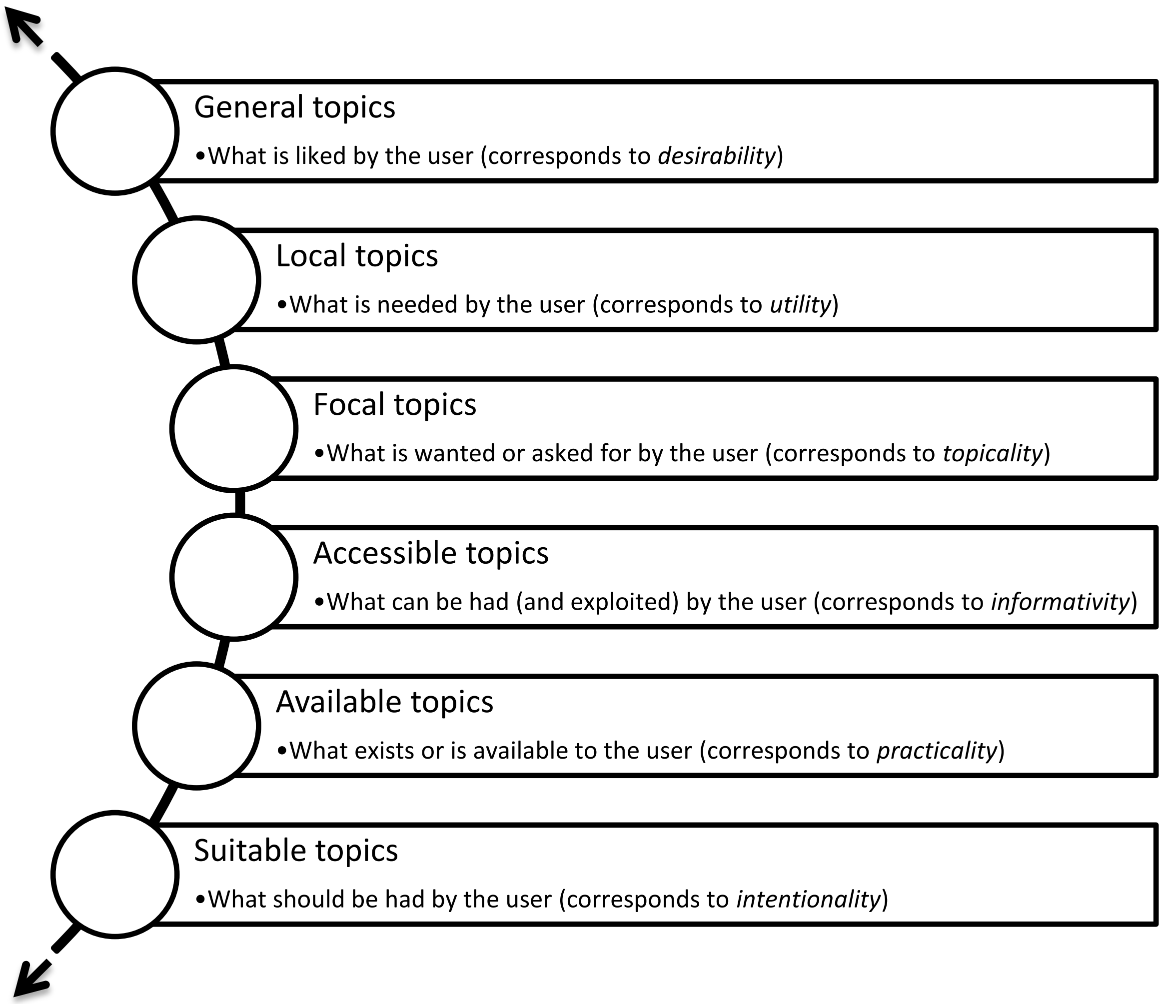


Fig. 6.3. The hierarchy of relevant contexts (HRC) revised. The order represents the sequence in which humans would most likely note the relevance of such topics, proceeding from the center to the periphery.

Note that this correlation between the HRC and the HRD does not in any way mean that the HRC can replace the HRD or vice versa. They are entirely separate hierarchies of different, but related, phenomena; the HRC classifies topics in relation to a person living in the world, whereas the HRD is a typology of resources as evaluated by this person. The correspondence between these two hierarchies may indeed be seen as the very essence of relevance. The relevance dimensions provide measures of how well a resource fits into the various HRC categories conjured up around users by their being in a situation in the world.

It may strike the reader that the order of the topic spheres as presented above differs from that of the relevance dimensions. This is because the HRC reflects the order in which topics become apparent and important to a human being in a situation in the world, starting from the most immediate and concrete and going towards the more abstract. Somewhere along the way the person's conscious awareness of relevance may even cease to proceed further. The HRD, on the other hand, reflects the order in which a resource is analysed and its relevance assessed. As pointed out by Sperber and Wilson (2002), human attention has evolved to focus on the most immediately relevant, and this is apparent from the HRC. With resource assessment, however, the order is different. Because in this case we are consciously selecting what is worth attending to rather than relying on our evolved abilities to sort stimuli on the spot, and because we are facing a number of potential stimuli that is so great, we start by narrowing down the amount of information by ignoring what is clearly irrelevant, i.e. it is suggested that we naturally proceed from the criteria that are most clearly associated with irrelevance assessment to those more weakly so associated, and then do the same with the criteria associated with positive relevance assessment. This accounts for the order of application of relevance criteria in search, and hence for the arrangement of dimensions presented in the HRD.

6.3 Shades of grey – Degrees of relevance

Relevance is recursive; a resource may be relevant due to its being relevant to something relevant. This was in fact the basic idea behind inferential systems such as the one described by Cooper (1971). However, the strength of such chains of relevance may quickly wear off because of the speedily growing processing efforts required[131] and the decreasing relevance of the information to the original topic. Hence, a resource may in principle be either relevant or not, but at the same time more or less **directly relevant**. A LAT should find the most directly relevant resources with regard to the query. This suggestion mirrors the idea from Relevance Theory that relevance is judged as the amount of gains obtained from a resource as weighed against the costs in processing efforts gone into extracting those benefits (Wilson & Sperber, 2004). This measure shall here be accepted as basically correct, though rather superficial and non-explanatory. A much more complex view shall be adopted.

[131] This was not a problem for Cooper, because the task would be performed by the system. In other cases, the user (or other person) has to identify why something might be relevant.

For decades it has been fervently debated whether relevance is a scalar or a binary phenomenon. We shall take the view that neither position is satisfactory on its own. Most resources have some value in between the two extremes of *relevant* and *irrelevant*, but even so the relevance as judged by a user may be binary because he or she may (consciously or subconsciously) have decided on a level that the value must exceed in order for the resource to be considered relevant to him or her in that instance. It is noticeable, however, that the resources that do pass this test seem still to be evaluated in a scalar manner so that some resources are considered more relevant than others.

6.3.1 The bisubjectivity of relevance

Relevance is an analogue phenomenon; a resource can be relevant to a certain degree. It can even be measured according to a multitude of different criteria. Even so, from the moment we observe it we can only handle relevance as a digital phenomenon. This does not mean that we cannot appreciate that one resource may be more relevant than another without any of them being irrelevant. It means that apart from binary judgments of relevant and irrelevant, we can only compare resources among themselves; we have no way of conceptualizing the strength of relevance, but we recognize a difference in strength when we see it. We can only say of a resource that it is merely slightly relevant because we compare it to a hypothetical ideal (or at least more relevant) resource in our minds.

This dichotomy between the analogue phenomenon of relevance and the digital perception and use of it means that if we want to measure the degree of relevance, we need to decide whether to measure the analogue or the digital version. For reasons that will become clear, we shall call these two values **a-** and **b-relevance** (which by a stroke of mnemonic luck happen to correspond to the analog and binary version, respectively). This is a case of what we have termed *bisubjectivity*; the object studied has some subjectively defined value, and at the same time the subject decides a criterion of accept. While the terms *analogue*, *binary* and *digital* refer to the way the value is fixed, *a-* and *b-relevance* may be considered stages in the assessment process; first the relevance is assessed as having a certain analogue value, and then it is assigned a b-relevance value in context. A resource may well be *a-relevant* to a certain extent only to be deemed *b-irrelevant* because of high thresholds of requirement. For instance, a resource may be highly topical, but if the user wants the answer to a specific question and the resource fails to provide it, it is not b-relevant.

A third kind of relevance may be introduced, which could be called **c-relevance**. If a-relevance is a measure of the actual usefulness of a stimulus and b-relevance is present when it is consciously deemed to be relevant, c-relevance covers the understanding of the term that we have been somewhat unwilling to assign to the phenomenon at all above, the relationship between a query and a document.

This is a question of relevance by definition, through the representation, within a given system with its defined rules and algorithms; the objective type of relevance (section 4.3). These three types of relevance correspond to the three relations *need*, *request* and *search*, respectively, and thus seemingly exhaust the possibilities for types of relevance in this sense (fig. 6.4).

Relation	Type of relevance
Need	A-relevance: Actually needed; relevant according to the characteristics of the resource
Request	B-relevance: Consciously judged to be relevant; relevant according to the requirements of the user
Search	C-relevance: Relevant by definition; relevant according to its (linguistic) representation

Fig. 6.4. The three basic types of relevance.

The types do not in themselves represent three different phenomena each with their own interest, but rather various ways of measuring one and the same phenomenon – relevance. The interest here lies in the different relationship to the user's psychological stance; while c-relevance is completely devoid of any contact with the user's cognitive operations and as such is a very insufficient proxy for relevance in any real sense (depending of course strongly on the degree of sophistication of the system), b-relevance is entirely dependent on the user's conscious and rational views – which may provide for immediate satisfaction but not necessarily ultimate success – and a-relevance represents the actual value of a stimulus as regards its potential positive cognitive effects on the user. The latter may not initially be appreciated by users, but in the long run should be the best help they can be given. What choice the designer makes is entirely dependent on the kind of system that one intends to produce.

6.3.2 Relevance scales

A-relevance can be measured according to the following three scales.

- **Strength**

  The first scale concerns the degree to which the resource is relevant (a-relevance) and how much the criterion in question counts (b-relevance). This is a qualitative measure of gains. Assigning weights to different criteria is crucial to an IR system as different people value criteria differently. The system should try to establish how desirable, useful, (directly) topical, informative, practical and intentional a resource is according to the user. Measuring actual satisfaction or utility values is very hard, and measuring the strength of topicality is notoriously difficult as similarity is not easily quantified (Schamber, Eisenberg, & Nilan, 1990; Aime, 2011). It should be noted that the strength of topicality is not necessarily equal to the processing efforts required to identify whether two topics are related, though these measures are likely to be often correlated. Approximate measures of informativity, practicality and intentionality would not seem to be unattainable, though certainly not easy to achieve either.

- **Amount**

How much information does the resource provide? This is a quantitative measure of gains. Being a set of statements, a document may contain a greater or smaller number of relevant sentences. Some parts of the resource may appeal while others do not, only one part may be on the right topic, a part may be written in another language, etc. The absolute amount of relevant material is more important than the proportion of a resource that is relevant. Thus, this value should not be considered a fraction but an amount.

- **Ease**

How easily could the gains be obtained, and how much effort does it require of the user to work out the relevance of a resource according to a given criterion? This is a practical measure of gains. The measure applies to the meta-level, i.e. to assessing whether a resource is relevant, and not only to the efforts that go into actually interpreting the contents. However, reading a document is in fact how the user judges whether it is informative, and hence the efforts are likely to be the same. The fact that it operates on the meta-level just means that efforts can also be spent in other dimensions, such as in order to determine whether a resource is topically relevant when it is somewhat far-fetched, or whether it is useful. Ease is the inverse of processing efforts spent on any of the relevance dimensions.

B-relevance comprises binary evaluation against predetermined requirements, and also digital comparisons of (real or hypothetical) resources. The analogue scales of a-relevance are subjectively interpreted by the user, but the values pertain to resources. Binary relevance consists in comparing these values to a pre-set requirement limit (the *set point* in the cybernetic model). If a resource fails to rise above the limit, it is considered irrelevant. Each scale has its own requirement value.

- **Required strength**

How a-relevant does the user require the resource to be in order to consider it b-relevant?

- **Required amount**

  How much information does the user require from a resource in order to consider it b-relevant?

- **Required ease**

  How easy does a user require it to be to obtain the benefits associated with accessing a resource in order for it to be considered b-relevant? Ease refers to all costs involved, such as processing efforts as well as time and money expenditures, and even induced psychological strain such as might be involved if the topic (or the way it is presented, including pictures) is unpleasant, or by the presence of too many ads, cf. the response returned by a participant in the study performed by Lopatovska and Mokros (2008) as a reason for rejecting a potentially relevant resource *too many pop-ups and ads*.

These limits are all absolute, though usually not fixed on a precise value[132]. However, the criteria may depend on each other, giving rise to comparative evaluation of criteria. Thus, the required amount may depend on the value of strength; if a resource is only weakly relevant, the user may nonetheless consider it relevant because of the sheer volume of information it provides. Likewise, required ease (i.e., tolerated costs) may depend on the amount of information obtained and the strength of its a-relevance. This, then, is how the supposedly basic evaluation of gains vs. costs suggested by Relevance Theory arises. The criterion of ease (cost) is used as a dynamic limit for binary judgment of the amount and strength of relevant material. This is hardly basic, given that it conflates all dimensions and ignores the difference between amount and strength, but it does capture the ultimate judgment of b-relevance. With the three scales in place, we can see that there are other oppositions that merit discussion apart from the one focused on by Relevance Theory:

- I don't get much information, but on the other hand it is fast and easy to read (amount low but ease high)
- I don't get much information, but what there is of it is highly relevant (amount low but strength high)
- I don't get very strongly (directly) relevant information, but on the other hand there is a large amount of it (strength low but amount high)

[132] Exceptions to this would be absolute requirements of text length, time limits, and limited funds for the purchase of resources.

- I don’t get very strongly (directly) relevant information, but at least it is easy to read (strength low but ease high)
- I don’t get information that is easy to read, but on the other hand it is highly relevant (ease low but strength high)
- I don’t get information that is easy to read, but on the other hand there is much of it (ease low but amount high)
- I don’t get much information, but what there is of it is highly relevant, and it is fast and easy to read (amount low but strength and ease high)
- I don’t get very strongly (directly) relevant information, but on the other hand there is a large amount of it and it is easy to read (strength low but amount and ease high)
- I don’t get information that is fast and easy to read, but on the other hand it is highly relevant and there is a lot of it (ease low but strength and amount high)

Surely, all of these relative evaluations should be mentioned in discussions of relevance and not only the last one, the combination of amount and strength versus ease?

Note also that each of the analogue criteria can only be measured while holding the others constant. Thus, measuring the amount of a-relevant information (according to any dimension) presupposes a binary decision on what is b-relevant, presumably in terms of strength. In other words, stating that a resource contains a certain amount of topically relevant information makes no sense unless it is also established what strength of topicality is meant. After all, a given resource may contain a large amount of peripherally topical information but only one sentence that is highly so. Also, the ease of reading a text (in the sense of limited costs) depends on how much of the text the user intends to read, so measuring ease involves deciding how much of the resource to include in this evaluation. In conclusion, the measurement of relevance is suffused with bisubjectivity. Assessing the relevance of a resource is therefore not a mere matter of judging whether it is relevant or not, even with the six dimensions in place; it involves a cybernetic circle of deciding on criteria and limits and evaluating whether a certain resource is high enough on various scales to allow for the low level of others, while constantly updating the levels of requirement against which the evaluations are made. Obviously, since all relevance scales depend on keeping others constant it is simply not logically possible to reach an objective conclusion; any judgment will necessarily depend on the decisions made on other scales, and so relevance assessment is effectively circular. This is especially so if the user has strong demands, such as a limited patience with long texts or a high requirement towards relevance strength. If, on the other hand, the user is willing to read long texts then the amount is effectively equal to the length of the document and this provides a basis for assessment of strength and ease.

Before an illustrative example can be given, we need to look at the various scales of relevance gradation in the six dimensions.

- **VI – Desirability**

  Desirability strength is about how appealing the resource is. This is a general measure of the impression formed from going through all dimension stages. An initial impression will be formed immediately when the search is initiated, and will be modified continually throughout the process. Thus, desirability represents the overall impression and satisfaction rather than aesthetic appeal only.

  The amount of material that fulfils the user's criteria of satisfaction may also be important.

  Desirability ease refers to the amount of emotional strain involved in accessing the resource. If a document contains many unpleasant pictures or passages, or if it is simply boring, ease diminishes.

- **V – Utility**

  Utility strength refers to the extent to which the resource is useful for the purpose for which it was required and applicable to the task. It may turn out that highly topical information is nonetheless not applicable to the case at hand. Strength diminshes as a situation is further removed from the ideal situation in which the resource would have been truly helpful, but low strength may be tolerated if the amount of relatively useful material is correspondingly large.

  The utility amount refers to how much useful material is provided by the resource. A low amount may be tolerated if the resource is strongly a-relevant.

  Utility ease refers to how easy it is to apply the information to the situation and the task. High costs may be tolerated if the combined gains from amount and strength are high.

- **IV – Topicality**

  The strength of topicality is measured as the cognitive distance between (sub)topics. Two subtopics may be related (at least) through identity, partiality, causation, co-occurrence, or association. This is notoriously difficult to measure (Schamber, Eisenberg, & Nilan, 1990).

The amount of information that is on the topic may be crucial. If it is only distantly related to the topic, much more information may be required in order for the resource to be accepted as b-relevant.

Topicality ease concerns the ease with which users are able to incorporate the information in their existing stock of knowledge and understand the topics (as opposed to the texts). Low topicality ease may be tolerated if it is realized that the resource is indeed highly relevant to the topic. If a user suddenly understands the topic better, the resource may turn out to be highly relevant, yet it has taken a lot of effort to get there.

● III – Informativity

Informativity strength is about the extent to which the resource is capable of changing the user's understanding or affecting him or her cognitively in other ways. Low informativity may be caused by lack of information in the document, lack of understanding on the part of the user, lack of new information not previously known to the user, or by mistrust or dislike towards the author, the document or the subject, causing the user to refuse to accept the information.

A resource which is low in informativity for any of the above reasons may still be considered relevant if it compensates by containing large amounts of information. It may seem incongruous to talk of a resource with low informativity providing large amounts of information, but here again we might distinguish between a-information which is there in the resource for suitable users to extract and b-information which the user is able to extract. If there is a lot of a-information, the user may decide to keep the resource for further study and hopefully he or she will be able to extract more of it.

A text may be difficult to understand because it is not well expressed, is too advanced for the user's level of understanding or is directed at a different type of user who would be more likely to understand it. Very difficult resources will only be accepted if they provide substantial gains.

Informativity ease concerns the ease with which the user is able to extract the information. This is dependent on the clarity of the linguistic expressions but not directly on the difficulty of the topic, which belongs with topicality ease. However, in a topically very difficult text there may even be sentences that the reader does not understand.

Wilson and Sperber give the following example of varying levels of informativity ease (Wilson & Sperber, 2004, p. 252).

Mary, who dislikes most meat and is allergic to chicken, rings her dinner party host to find out what is on the menu. He could truly tell her any of three things:

1. We are serving meat.

2. We are serving chicken.

3. Either we are serving chicken or $(7^2 - 3)$ is not 46.

All of these answers would be correct and hence a-relevant. On the other hand, answer 2 would be more relevant than answer 1 because it is more informative, and it is more relevant than answer 3 for reasons of topicality ease; even though answer 2 and 3 are logically equivalent, answer 3 requires considerably more processing effort and hence is likely to be considered less b-relevant.

⬤ II – Practicality

Practicality strength encompasses the availability of a resource in a format and language readable by the user with a certain device, and the extent to which the presentation is clear and practical. If a resource cannot be accessed, or is in a language or script entirely unintelligible to the user, practicality is zero.

If the presentation is very bad or impractical, the information may only be accepted as relevant if there are large amounts of it. Consider a non-downloadable site with many pages and complex cross-referencing. In many cases, users may prefer a downloadable document, but if the site compensates for this by containing much larger amounts of information or a pleasing user-friendly design, it may still be deemed at least as relevant.

Practicality ease represents the ease with which a document can be imported through the user's device, including costs in time, money and effort. With practicality, strength and ease are seemingly closely related. Strength represents the difference between the current system and one that can handle the resource in question. Very impractical resources will only be accepted if they provide a lot of material. For instance, it is not inconceivable that the existence of a large amount of highly topical information on a very specific topic would induce a user to acquire special software required to open the document and even learn Japanese, if this were necessary in order to access the information. However, this would not be likely if the amount and strength of relevant information were more limited.

● I – Intentionality

Intentionality strength refers to the similarity of the user to the kind of reader for whom the resource was intended. The user may decide that he or she only wants information that was at least intended to inform similar people in similar situations because that would make the information directly applicable to the task at hand.

If intentionality strength is weak, the user may decide that more material is required than had it been intended more specifically for a person in that kind of situation. This is because more material may be necessary if the user is to find something relevant in that particular situation.

The amount of information intended to be read by the user is perhaps not a criterion used very frequently in the final assessment. It is, however, the only amount value available in the earliest stage of the search process. A low amount is tolerated if it was strongly intended to be read by that particular user in the situation in question.

Intentionality ease refers to the social benefits of reading a certain resource. In other words, this value is equal to the social cost of *not* reading it.

Here is the example: Imagine a user looking for information on a certain difficult topic. He starts by looking for highly topical information in the resource he is currently assessing. He finds that there is very little that looks as if it is exactly right, but he then lowers his expectations as regards strength, which causes the amount of information passing this new threshold to rise. It also lowers ease, because by now there is more information and so it will take more time to read. Seeing that there is now a substantial amount of reasonably topical information, he may decide that it is a relevant document, but this final decision may depend on the availability of other resources with higher levels of topical strength. Given that the topic is a difficult one there is not much he can do about the fact that the topical ease is very low; he will simply have to accept that. On the other hand, this may cause him to prefer short texts and to avoid texts that are only peripherally topical. Consequently, he may prefer another resource. Having found a resource with acceptable relative levels of topical strength, amount and ease, he looks for useful passages. Imagine that what he actually needs is a quote. Even if a resource is strongly topical it may not provide a useful quote that is applicable in the kind of situation for which the user needs it. The search is driven by the need for something that is useful in that particular situation, so the user has a high demand as regards utility strength and it does not do to lower the threshold in this case. Consequently, he must accept lower levels of amount and/or ease instead. Maybe he can find a shorter quote than he had hoped for, or one that is difficult and hence will require more effort to understand (and even find). He may also have to read more text than he had hoped in order to find the quote, which will lower ease further. In fact, he may have to lower the threshold of utility strength temporarily in spite of the demand for strong utility in order to find the quote, thus increasing the amount of information that he is required to read (this effectively turns the task into one of finding a quote rather than applying it to the external situation of need, for the time being, which shows us that the process may contain successively embedded tasks). He will now have to decide whether this is worth doing, depending on the likelihood of finding other resources where the amount can be kept low. Having read some of the text he may decide that it is simply too boring, or the topic is too unpleasant, so he will have to find a different source, unless he can keep the amount of information that he must read to a minimum. On the other hand, he may be reconciled to the fact that texts on this topic are going to be boring, which means that he has low requirements to desirability strength. Even so, he may be squeamish when it comes to the unsavoury illustrations, thus demanding a high desirability ease (accepting limited emotional strain). This will then lower the acceptable amount of information that he can bear to read.

Before getting this far, the user has made certain unspoken demands as regards intentionality, practicality and informativity, whether knowingly or not. Imagine that he needs the text to be scientific and will not accept popular scientific publications, i.e. he has a relatively high threshold of intentionality strength. Imagine that it turns out that he has to pay in order to access the document. He may immediately decide to look for another resource, if he has a practicality ease requirement that says no money should be spent, or he may compare the price to the amount of information he will get. Ideally, of course, he would have to know and take into account what the overall relevance of the text is in order to be able to judge whether the price is reasonable. Assuming that he finds a resource, he still has to be able to read and understand the text, and thus informativity strength requirements are necessarily set reasonably high. Given that it is a difficult topic, he may of course be prepared to find that he does not even understand all the sentences, and hence he may have to lower his requirements. This will increase the amount of text that he is willing to read, with reasonable but limited confidence. This is then the amount of text that is the input to the topicality considerations discussed at the beginning of the example.

6.3.3 Trust and appeal

Any of the six facets of a resource may be disliked or distrusted, leading to decreased desirability, but also secondarily to reduced informativity. A single feature of a resource may thus affect several dimensions, and this is one of the phenomena which may seem confusing about relevance when asking people about their actual assessment criteria in concrete cases.

The intentional resource may be disliked if the user feels that he must read a text and this sense of being forced in itself lowers the appeal of reading it, or if for some reason he feels uneasy about the way he obtained the resource or came to choose it rather than others. It may be distrusted if the reader is unsure about the honesty or sincerity associated with the author's intention to inform.

The substantial resource may be disliked for its physical properties and qualities, and it may be distrusted if there is a risk that it is infected with a computer virus, for instance.

The formal resource may be disliked for its linguistic or other aesthetic qualities beyond mere handling, and it can only be trusted if the user is confident that the author has been able to communicate his true meaning, which may not be the case if he is not using his native language for instance. This would be a distrust in the words rather than in the author, because it is uncertain whether the words really convey what the author meant.

The interpersonal resource may be disliked if the contents themselves are boring or otherwise unappealing, and distrusted if the user has reason to doubt the validity of the information, i.e. whether it represents "the truth".

The effective resource may be disliked if it has unpleasant psychological effects, or if reading it or choosing it as a source is deemed to result in negative consequences of any kind in future. It may be distrusted if the reader feels manipulated when reading or experiencing the resource.

On the other hand, each facet of the resource may of course equally be particularly liked or trusted, so appeal and trust may contribute in both directions. This is the case, for instance, when one participant in a study of real-life relevance assessment criteria reports that *"the author is my major professor, so I should really look at this"* (Barry, 1994). There is no doubt, however, that when it comes to trust and appeal the negative characterizations are the more prevalent in such studies.

It is clear that all facets may be judged separately for desirability, and this in fact holds for all dimensions; any one of the six facets may be desirable, useful, topical, informative, practical and intentional. This means that relevance assessment is carried out on all of the six facets in parallel. Consequently, there are hundreds of possible relevance values which might potentially be measured. Even if we did have access to the user's mental processes, monitoring and modelling the assessment process would still be a major undertaking.

6.4 Criteria of relevance

When users assess the situational relevance of a resource they employ an array of conscious and/or subconscious relevance criteria, and before applying the Regulated Flux model to the data from behavioural studies we shall have a brief look at the kinds of relevance criteria that people report using. Barry (1994) provides an in-depth study of such preferential variables, which she groups roughly into those concerned with the following.

Barry (1994)

- The information contents of documents
- The user's previous experience and background
- The user's beliefs and preferences
- Other information and sources within the information environment
- The sources of documents
- The document as a physical entity
- The user's situation

It was pointed out above that people belonging to different domains value relevance criteria differently. Specifically, criteria such as recency and subjective or objective validity may be emphasized in a scientific community, where a much higher level of required processing efforts is also tolerated than elsewhere. Domains are thus characterized by specific preferences as regards the features of resources. In Barry's study, all respondents mentioned criteria of the above types for their relevance evaluation of actual documents. The study was carried out in a scientific environment, the documents being scientific papers. This is probably a domain where such criteria are at their strongest, but even so, they would be expected to be of general importance. Schamber (1991) is a similar study on weather information, using categories similar to those employed by Barry. Barry and Schamber (1998) confirm that even though there are clearly differences in the exact criteria employed there is a great deal of overlap and criteria can probably be set up on a more general level. With such a purpose in mind, we shall develop a typology that encompasses all the various kinds of criteria in a generalized framework.

Assessment experiments are normally carried out on informational resources only[133], and furthermore it is generally assumed that the information is of central interest with the substantial resource being secondary. Below, an attempt has been made at generalizing the terminology so as to include all kinds of responses, such as products that the user might be interested in buying, and in principle even actions carried out by the system. However, informational documents remain central to relevance studies in IR.

6.4.1 A typology of relevance criteria

Given that user opinions may either concern the substantial resource, the linguistic content or the source whence it comes, assessment criteria are here assigned to three different types according to whether the item assessed is the **content** information, the substantial **resource**, or the **source**. This is a practical, simplified distinction conflating some of the facets of resources. In the case of documents, the *content* is the formal, representational and interpersonal document, the *resource* (in the narrow sense used here) is the substantial document, and the *source* is the author (as well as others involved in the publishing, such as a journal). In the case of products, the *content* is the product viewed as an item with a function, the *resource* is the product viewed as a physical object with properties, and the *source* is the supplier (and others involved in the transaction). In other words, in this simplified version documents are basically treated in the same way as any non-linguistic product.

Assessment criteria are tentatively regrouped into the four generalized types explained below. It is felt that this, in combination with the content-resource-source dimension and the principle of bisubjectivity, explains the data better than earlier attempts at producing such typologies.

- **Novelty and interest**
  Novelty and interest measures concern the degree to which a certain resource is new to the user and provides interesting material of value.

- **Quality assessment**
  Quality assessments are about the actual value of the resource. Source quality assessments concern the reliability of the source and the author. Such criteria are prominent in the selection of scientific sources (Park T. K., 1993), but they are of course equally relevant in the choice of a web shop.

- **Applicability assessment**
  Assessments of the information's currency, geographic origin and level of detail are examples of applicability assessments; how likely is the information to be applicable to the user's

[133] However, Choi and Rasmussen (2002) is an example of an assessment experiment involving images.

situation? Such questions are crucial in connection with fact-delivering services such as providers of meteorological information (Barry & Schamber, 1998). Again, it is a kind of criterion often used by customers in web shops. If you want to buy a DVD, finding out whether it can be watched in your PAL region is an instance of applicability assessment.

● **Accessibility assessment**

On the other hand, deciding whether the shipment from the other side of the globe is too costly is a question of accessibility. Accessibility criteria concern the availability of the source, in a language understood by the user, delivered in a practical format. Accessibility also refers to the readability of the material compared to the user's level of expertise in the subject, as well as the way it is presented; layout and interactive features also belong here.

All assessments can be **user-based** or **resource-based**. A clear example is accessibility, which may be high when it comes to layout and presentation (resource-based accessibility), but low in a specific situation (user-based accessibility), because the user happens to have too little knowledge of the subject to be able to understand the text (or use the product, as the case may be). This is again an instance of bisubjectivity; accessibility clearly depends on features of the resource as well as on user capabilities.

More precise descriptions of the various criterion classes proposed can be found in the typology below, which represents an attempt at classifying the criteria reported in Barry's study, as well as those found in Schamber's follow-up study, into the types suggested here. Many of the reported criteria in fact conflate several types; for instance, affectivity implies emotional reactions to either the content or the source (or possibly the resource, though this is not mentioned by Barry). Also, novelty may trigger interest, as is clear from the responses Barry recorded from participants: *"I had never heard of these authors, so I would want to follow up on them"*; *"I would be interested, not just in this article, but in looking at this journal that I've never heard of"* (Barry, 1994, p. 155). Several additional criteria have been added in the typology, not least in order to exemplify assessment of product relevance, which was not covered in the studies by Barry and Schamber.

Typology of relevance criteria

1. Content assessment

    a. Novelty and interest

        i. **Content novelty**: This covers whether the user already knows the contents of the resource, or at least has access to sufficient alternative information. (*Content novelty, availability within the environment, personal availability*)

        ii. **Content interest**: This refers to whether the contents are likely to be interesting. If the contents are obvious and uncontended, they are unlikely to provide much interest. Also, information may be empty if it is simply stated that nothing is known or that something is wrong. This is not necessarily uninformative, but may nevertheless reduce interest on the part of the reader. (*Consensus within the field, external verification, effectiveness, depth, summary/interpretation*)

    b. Quality assessments

        i. **User-based content quality**: This is about whether the user likes or agrees with the information provided in the resource. (*Subjective accuracy/validity, affectiveness, source agreement*)

        ii. **Resource-based content quality**: This refers to whether the contents are correct and of high quality. With products, it also refers to the usefulness of the item. (*Objective accuracy/validity, effectiveness, tangibility, depth, specificity*)

c. Applicability assessments

   i. **User-based content applicability**: This refers to whether the contents are applicable to the user's background and current knowledge. (*Experience/background*)

   ii. **Resource-based content applicability**: This refers to whether the contents are applicable to the situation of use. With products, it includes size considerations, etc. (*Scope, recency/currency, geographic proximity*)

d. Accessibility assessments

   i. **User-based content accessibility**: This is about whether the user is capable of understanding or using the resource. (*Ability to understand*)

   ii. **Resource-based content accessibility**: This refers to whether the contents are presented in an accessible manner. (*Verbal clarity*)

2. Resource assessment

   a. Novelty and interest

      i. **Resource novelty**: This covers whether the user already knows or even possesses the substantial resource in question. (*Stimulus document novelty*)

      ii. **Resource interest**: This refers to whether the substantial resource is likely to be entertaining, aesthetically appealing or interesting in other ways. This is especially important in product search.

b. Quality assessments

   i. **User-based resource quality**: This is about whether the user likes the look and substantial characteristics of the resource, including the interface of websites. (*Dynamism, presentation quality*)

   ii. **Resource-based resource quality**: This may refer to objective considerations of colour and form, as well as length of a document. (*Variety/volume, depth/scope*)

c. Applicability assessments

   i. **User-based resource accessibility**: This is about whether the substantial resource is of a type that will make the contents accessible to the user, such as the format of a document.

   ii. **Resource-based resource accessibility**: This refers to whether the substantial resource is easily and cheaply available. (*Obtainability, cost, time constraints, visual clarity*)

3. Source assessment

   a. Novelty and interest

      i. **Source novelty**: This covers whether the user knew the author or the journal etc. beforehand. This may have either a positive or a negative effect on the relevance of a resource[134]. (*Source novelty*)

      ii. **Source interest**: This refers to whether the author or source is of special interest to the user. (*Relationship with author*)

[134] This may in fact be true of all kinds of novelty.

b. Quality assessments

i. **User-based source quality**: This is about the user's emotional reaction to the author or source. (*Affectiveness*)

ii. **Resource-based source quality**: This refers to whether the user trusts the author, source or seller. (*Source quality, source reputation/visibility*)

I would like to emphasize that this typology does not represent the understanding of relevance put forward here. Rather, it is a suggested classification of relevance assessment criteria. Because these are only indirectly related to actual relevance phenomena, it makes sense to have a classification of criteria which is separate from the actual relevance model. It is clear even from the list of groupings used by Barry that these belong to different dimensions and facets and are not at all of the same type. As noticed by Xu and Chen (2006, p. 962), almost all studies of assessment criteria have been "exclusively exploratory and data driven", and consequently unable to identify the foundational principles behind the many suggested criteria and the relative importance of the various criteria[135]. In contrast, the approach suggested here attempts to uncover the roots of relevance and assign the surface criteria identified in exploratory studies to different underlying processes.

6.4.2 First- and second-order criteria

Relevance is a phenomenon that comes naturally to everybody; everyone knows what it is and how to apply this knowledge in assessing stimuli, yet few people actually understand what is going on in a deeper sense (Saracevic, 1975). Consequently, the replies made by test subjects in user studies should not necessarily be taken to answer the question posed; many of the criteria reported on in the literature do not describe features that make a document relevant but rather explain the presence of those features in the first place. For instance, rather than pointing out the capability of a document to change the reader's understanding of the topic it treats, it comes much more naturally to users to explain that it is written in a clear style. Yet, the clarity with which it is written does not in itself make a document relevant. It does, however, enable the reader to more easily understand the information

[135] Bruce (1994) claims that it is possible to measure such strengths of criteria and makes an attempt at doing so. However, he uses a small number of criteria which are very particular to the context that he is investigating. It would be interesting to test the general types of criteria described above in a similar manner. Lopatovska and Mokros (2008) provide some interesting results in that direction.

provided in it, which in turn increases the relevance of reading it. Subjects may also point out that they know the author of a document personally, thereby implying that this makes it relevant to them, when in fact what makes it relevant is that their reading the document may give them social advantages in relation to the author, or spare them from the opposite, in addition to any practical benefits gained from following the author's work[136]. Alternatively, the user's desire to read the author's contributions may simply be due to a supposition that they are likely to be topically highly relevant, or even pleasant reading. The criterion as such does not, therefore, reflect the underlying causes for judging a resource to be relevant.

By way of criterion analysis, study authors such as Barry (1994), Barry and Schamber (1998), Lopatovska and Mokros (2008), and Choi and Rasmussen (2002) group responses that refer to similar features in a way reminiscent of the typology presented above. These groups do not have anything to do with the kind of relevance that they represent, but then it would not be possible to group them unambiguously in such a way since one second-order reason may well affect several relevance dimensions simultaneously. Frequently, one second-order criterion explaining the presence of a feature may encompass several first-order criteria, as one feature may cause various simultaneous advantages or disadvantages to the user. For instance, the clarity and layout of a document may influence its appeal to the user as well as his or her ability to understand the contents. These are entirely separate first-order criteria of relevance, yet they are subsumed under the same criterion in user studies because they derive from the same surface features of the resource. Describing the characteristics of relevant resources does not lead to an adequate understanding of the phenomenon of relevance; the criteria need to be thoroughly analysed. Finally, having such a large number of second-order criteria makes it impossible to predict further criteria or establish the relative importance of criteria. Yet both are exceedingly important for a theory of relevance to be useful (Bruner, 1986). It is clearly necessary to abstract from the concrete explanations provided by test subjects and attempt to disentangle the criteria in order to get at the true essence of relevance. This we have done above; it now remains to reconnect the resultant model to the empirical data in the form of reported relevance assessment criteria.

6.4.3 Testing the model

The following are criteria mentioned by participants in the empirical studies. While the authors of the original studies can only verify the use of these criteria and discuss why this might be the case, the

[136] This is a criterion that at least plays a role in a scientific domain, cf. *"The author is a friend and an ally, so I want to see what he's doing"*, and *"The author is my major professor, so I should really look at this"* – responses provided by participants in the study conducted by Barry (1994).

Regulated Flux model suggested here is able to explain why these criteria are the ones that matter and place them in a larger perspective.

One criterion frequently mentioned in the literature is *novelty*. This in fact covers several different phenomena relating to whether a document or the information it contains (or even an author or journal) was previously known to the user. Novelty is not a simple phenomenon and does not in itself lead to relevance. Content and document novelty first and foremost concern informativeness; if the contents are already known, the reader may be assumed to gain nothing new from attending to them. However, new information and new resources may also appeal to the reader as being more interesting than what is already known. This is true of *stimulus document novelty* and *content novelty* as well as *source novelty*. Hence, novelty affects desirability. *Availability within the environment* is defined by Barry (1994, p. 155) as "the extent to which information provided by the stimulus document is available through other sources". Presumably, the users in question have already read such alternative resources. Hence, the resource in question is not likely to provide much new information and this is the reason for their rejection of the new document on the same topic. The user may even possess similar information in his or her collection already (*personal availability*). If the user already has a copy of the *same* resource, this is a case of lacking *stimulus document novelty*. Note that the present model treats the negative effects of lacking stimulus document novelty differently from the positive effects of discovering a new document that was not previously known to exist: *"This is an article he's written that I didn't know about"* is a statement that implies a heightened interest (i.e. desirability), whereas *"I already have this article, so I wouldn't get it"* indicates a lack of informativity. *Consensus within the field* may likewise have an effect on interest (desirability), cf. *"This is a subject that's very much up for debate"*. In such cases, the fact that there is no consensus is not what makes the resource relevant; rather, it spurs interest, which in turn makes it relevant. However, the very fact that there is disagreement may indeed sometimes, at least in a research context, in itself make the resource useful, by defining a task (viz. resolving the disagreement), cf. *"There is little debate over this issue of public policy for handicapped groups, so this would not be as useful to me"*. An interesting criterion which can be assigned to desirability is *contrast*, mentioned by Maglaughlin and Sonnenwald (2002), cf. *"That might make an interesting comparative"*. In this case, the user appears to choose to access the information exactly because it is not presented from the anticipated perspective. *External verification* is also not a criterion that in itself makes a resource relevant; rather, it stimulates trust in the author and the information provided, which in turn enables informativity. *Effectiveness* may sometimes behave similarly, but a negative result such as a method that is shown to be ineffective may also make the resource less interesting to read, and of course the method is less useful, and this may be considered to be true of the text as well. *Objective accuracy/validity* and *tangibility* are also mostly about trust. If the user knows some of the information to be incorrect, he or she will lose confidence in the author and content. Also, tangible and tested information may increase confidence in its veracity. *Source quality* and *source reputation/visibility* are about trust in the source, such as a journal. This affects informativity. *Ability to understand* and *Experience/background* likewise involve informativity; coming from a different background may seriously weaken the user's understanding of a text. With *experience/background*, the information may also not be applicable to the task since it is treated differently from the practice in the user's community. *Verbal clarity* is another informativity criterion. If the text is difficult to read, it is less likely that the user will understand it; it will certainly require more effort. Thus, verbal clarity is mostly measured along the scale of ease. It is also likely in many cases to reduce desirability considerably: *"Now we come into a morass here where he starts repeating himself, and not very well".* One of the responses shows a case of predicted verbal clarity: *"Most documents from that publisher are well written and edited"*. This is an example of an opinion about a relevance dimension long before the final judgment is made. *Variety/volume* is about the sheer amount of information that the resource provides. *Depth* affects the strength and amount measures of informativeness of a resource in that a superficial account may not provide enough details for the anticipated gains to be realized. *Scope*, *recency/currency* and *geographic proximity* are about the delimitation of the topic. A topic may be restricted in geographic and temporal extent as well as in the subjects with which it deals. *Relationship with author* refers to whether the author is of special interest to the user. Such a relationship may cause a certain amount of desirability, but more than anything it involves intentionality (and sometimes utility): The user feels he or she ought to read the document, because it is useful, practically or socially, to follow the work of the author. There may sometimes be a practicality component attached to this criterion as well in that it will be easy to get further information (or help) from the author if needed. *Subjective accuracy/validity*, *affectiveness* and *source agreement* are about the extent to which the user likes or agrees with the information provided in the resource. These criteria involve some desirability (cf. *"I simply enjoy reading articles about the footraces"*), but also informativeness, in that – as with lack of trust – people do not incorporate knowledge that they do not believe, and this may sometimes also be true if they simply do not agree with it or even like it. As with controversial topics, subjective accuracy/validity may sometimes make a resource useful in a scientific context: *"This is going in a direction I would argue, I could use it to bulwark my argument"*. Boredom, too, may inhibit informativeness: *"This is just an incredibly boring topic"*. *Affectiveness*, which is a rather broad category, may also involve a like or dislike towards an author or source. This will affect informativeness as well as desirability. *Dynamism* and *presentation quality* deal with whether the user likes the look and substantial characteristics of the resource, including the interface of websites. Again, there is some desirability involved, but at the same time a good presentation may stimulate informativity by making the information easier to absorb, cf. *"Hard to navigate"*, *"Summary, table of contents, drilldown are all great"*, *"The letters are too small"*, and *"I do not like some features of the design"*, all responses from the study by Lopatovska and Mokros. As noted by the authors, other criteria that belong here are the presence of links to other information sources in the text, and the use and quality of illustrations. *Obtainability*, *cost* and *time constraints* are practicality measures representing situations in which pre-set practicality constraints make some resources irrelevant.

Very few studies have looked at assessment criteria for other kinds of resources than text (Choi & Rasmussen, 2002). The criteria used in studies of image relevance evaluation differ substantially from the above in name, though not in sense. What is known as *topicality* in this connection refers to how the image relates to the user's task, i.e. it is really utility. *Accuracy* is more akin to topicality in the received sense and belongs in the dimension of that name. *Time frame* similarly belongs here. *Suggestiveness* and *novelty* are informativity criteria, as is *completeness*, which corresponds to depth (and probably also scope) in the text relevance studies. *Accessibility* is clearly a question of practicality (cf. *"This photo is less useful. But it is important because of the difficulty of finding other relevant pictures"*). *Image clarity and quality* is a combination of informativity and desirability. Finally, *appeal of information* and *technical attributes of images* concern desirability. The latter covers the mood or emotion of an image, point of view, and colour, and consequently relates to the rational rejection of resources that do not fit what the user had anticipated, in spite of being topical and even potentially useful.

All of the reported relevance criteria can be explained by the model proposed here, though they cannot always be easily assigned to a dimension as they are frequently complex and of a second order. It is not always easy to disentangle the effects on the various dimensions, because psychological effects often spread in intricate ways. One example of such a situation is the scenario provided by Blair (2006) in which a manager asks his or her assistant to *Get me any reports that analyze Central European investment prospects*.

> For example, is it a request for the investment records of specific investors, or is it a request for general investment trends? Is it a request for historical information combined with present results, or is it a request for an informed analysis of future prospects? Should the information be broken down by company, or industry, or by

> country? Should the information include just economic information/projections, or should there be an analysis of the political situations and economic stability of the concerned countries? Should the request only be directed toward internally done analysis, or should it be targeted towards gathering analysis done by those outside the organization? Are there free government sources of this kind of information—in which case, how would you find out where they are—or is this information maintained by commercial information brokers—in which case, how much do you want to pay to get it? Does the person who wants the information have a strong background in economics and finance? If so, the information can be very technical. If not, more general analysis is needed, etc. This list of variations can be extended indefinitely. (Blair, 2006, p. 87)

Most of these questions relate in the first hand to topicality, but with side-effects in various other dimensions. Let us break it down.

> is it a request for the investment records of specific investors, or is it a request for general investment trends?

This is a topicality matter, but also one of utility; the type of resource needed is the one that will help solve the task.

> Is it a request for historical information combined with present results, or is it a request for an informed analysis of future prospects?

This seems to be a purely topical question.

> Should the information be broken down by company, or industry, or by country?

This is an interesting question, because it deals with different angles on the same information, which is largely a matter of informativity as the presentation of the data is crucial for the ease with which the reader can extract the information needed. However, if the presentation choice is particularly unfortunate from the point of view of the reader with a specific use-situation in mind, it may even be a matter of practicality, if it becomes effectively impractical to extract the information because everything has to be recalculated in order to get the desired angle. The reader may simply not be willing to invest the required amount of effort in the task. Normally, effort issues pertain most strongly to the informativity dimension, but practicality may block informativity entirely if the reader cannot even get started because no access can be had to any useful information to extract.

Finally, the way something is presented is also a matter of taste and preference, and hence the question above also relates to desirability.

> Should the information include just economic information/projections, or should there be an analysis of the political situations and economic stability of the concerned countries? Should the request only be directed toward internally done analysis, or should it be targeted towards gathering analysis done by those outside the organization?

This is again topical, but with reference to utility, as is often the case.

> Are there free government sources of this kind of information—in which case, how would you find out where they are—or is this information maintained by commercial information brokers—in which case, how much do you want to pay to get it?

The only question relating to relevance here is the last one, which is a matter of practicality.

> Does the person who wants the information have a strong background in economics and finance? If so, the information can be very technical. If not, more general analysis is needed, etc.

This is clearly a matter of informativity. However, there can be aspects of utility and desirability involved as well.

The original user behaviour studies collected and presented long lists of seemingly unrelated responses. They grouped the responses intuitively, but did not explain them, and the results did not enable the authors to predict unattested criteria (Hjørland, 2010). As noted by Xu and Chen (2006, p. 961), "having a large number of factors obscures the key factors". Xu and Chen reduced the number to five core relevance criteria: *Scope*, *novelty*, *topicality*, *reliability* and *understandability*. These account for many but by no means all of the above criteria. The present model has ordered and explained all of the criteria. Furthermore, it is indeed possible to predict relevance criteria that did not appear in the studies on the basis of this over-arching model. Future work should test the model proposed here in a manner similar to the one suggested by Xu and Chen. They test their own limited model and find that it accounts for 52% of the relevance variation observed in their study. It is expected that the model presented here will in principle account for all non-random variation, at least if any unforeseen criteria that can easily be assigned to one of the six dimensions are allowed to be included as part of the model, which is exactly what the Regulated Flux model allows.

6.4.4 A definition of relevance

People who like definitions will be longing for one at this point. However, a definition of relevance may not contribute much to the understanding of the phenomenon. In a communicative context, Leech (1993, p. 94) defines *situational relevance* in the following way.

> An utterance *U* is relevant to a speech situation if *U* can be interpreted as contributing to the conversational goal(s) of speaker or hearer.

It is interesting to note that according to Leech an utterance *is* relevant if it *can be interpreted as* contributing something. This statement appears to confound potential and actual relevance; it seems more likely that an utterance is (or may be claimed to be) relevant if it *is* interpreted by someone as contributing something. Otherwise, it is merely potentially relevant. It may be that the hedge "can be interpreted as" was added in order to allow for cases in which people's interpretations differ. In such cases, the utterance can certainly be claimed *not to be irrelevant* if at least one person interprets it as contributing something. The interesting observation that can be made from this is that apparently everybody (who is relevant!) has to agree for an utterance to be irrelevant, but not for it to be relevant[137]; irrelevance is public whereas relevance is private and personal. Thus, the above definition needs to be replaced by two different definitions.

> An utterance *U* is relevant to a speech situation if *U* is interpreted by someone as contributing to the conversational goal(s) of speaker or hearer.

> An utterance *U* is irrelevant to a speech situation if *U* cannot be interpreted by anyone present in the speech situation as contributing to the conversational goal(s) of speaker or hearer.

[137] Of course, this is a feature of the definition, and the reader may disagree.

In IR, a state-of-the-art definition of relevance is provided by Hjørland and Christensen (2002, p. 964) as follows.

> Something (A) is relevant to a task (T) if it increases the likelihood of accomplishing the goal (G) which is implied by T.

This definition places great emphasis on utility and overlooks the other dimensions to a wide extent. Of course, in a way it only excludes desirability, because (so the Regulated Flux model tells us) the other dimensions precede the evaluation of utility and may thus be claimed to be presupposed by such a definition, but even so it simplifies things greatly.

Xu and Chen (2006, p. 962) adopt a much less utilitarian viewpoint with more emphasis on cognitive factors; their definition is as follows.

> Relevance is the perceived cognitive and pragmatic impact of the content of a document in relation to the user's problem at hand.

This is intended to include the user's understanding and use of document information in a situation. At the same time, the task on which utility assessment is based is also included.

In light of the multidimensional understanding developed above, we shall settle on the following definition.

> A stimulus is relevant to a person if that person thinks he or she will gain sufficiently from attending to it, and that this gain outweighs the investments associated with it.

This definition is more psychological than the task-based one proposed by Hjørland and Christensen. It is a cognitive standpoint which looks superficially very much like the one advocated by Relevance Theory, but being couched in the theory surrounding the Regulated Flux model the gain is to be understood in a multidimensional way which involves all of the relevance dimensions. No explicit mention is made in this definition of the situation, in spite of persistent claims that the situation is paramount. This is because any person is necessarily in a situation, and relevance is not restricted to the focal or local sphere of interest but applies across the board. It is consequently not something that is triggered in a situation, but rather in a life-world including the focal, local and general spheres perceived by that individual at that point in time. All of this has simply been expressed in the definition by using the word *person*. The difference from Relevance Theory thus appears not so much in the definition of the phenomenon as in the Regulated Flux model to which it is intimately connected.

The fact that the definition is so similar to that of Relevance Theory hints at applicability to discourse as well as search. However, it must be emphasized that it is *merely* a definition; describing when such a situation obtains is quite another matter. Without an understanding of the interplay between dimensions a definition says very little. Relevance cannot really be understood without digging deeper than a superficial definition which merely mirrors the final results of relevance evaluation. Many factors contribute to or limit the gains, so that estimating the potential benefits provided to a user by a given resource is no simple matter. Schamber *et al.* end their 1990 review by stating that they "consider the pursuit of a definition of relevance to be among the most exciting and central challenges of information science, one whose solution will carry us into the 21st century" (Schamber, Eisenberg, & Nilan, 1990, p. 774). Perhaps a definition is not what we are looking for after all; what we need is a model that illustrates and explains the processes involved in the phenomenon. The model in its entirety is a better definition of the phenomenon than a superficial statement. It is hoped that with the Regulated Flux model a better understanding of the notion has been achieved so that IR researchers may now concentrate on ways of satisfying the complex relevance requirements of users. After all, this may turn out to be an even more demanding task than previously acknowledged.

6.5 Search and communication

The notion of relevance has occasioned a continuous and heated debate in the IR community since the middle of the 20th century and countless types of relevance have been suggested through the years, supplanting and complementing each other. There can be no doubt that this is an intensely important concept of unsurpassed significance for most sciences. Nevertheless, and in spite of being duly fascinated by the concept, Saracevic (1996) feels that Schütz' theory exaggerates the importance of relevance in our everyday lives.

> The strength of this theory lies in explication first of the existence and then the interactivity and interdependence between various types of relevance. This is a powerful and useful insight. The weakness is in its breadth - it tries to explain all our actions and connections in the 'lifeworld' through relevance. For some, relevance is clearly irrelevant. (Saracevic, 1996)

Unlike Saracevic, I agree with Schütz and the proponents of Relevance Theory that relevance does indeed guide our behaviour in general. The problem is in recognizing the surprisingly wide scope of the effects of relevance. Since the Regulated Flux model encompasses appeal, an act may be said to be relevant simply because the agent thought it appealing to perform it. Certainly, the act may be characterized as completely irrelevant along other dimensions, but that does not remove it from the province of relevance. It is indeed hard to think of any act which would not be carried out with some reference to its relevance along at least one of the six dimensions of this fascinating phenomenon. Thus, the challenge is not to develop a model of less wide application but to accept the pervasiveness of relevance phenomena, even in situations where we are wont to conceive of the effects as being due to other causes such as appeal.

6.5.1 The role of relevance in IR and beyond

Whatever notion of relevance one adopts, we have seen that it is generally understood to be a relation between two entities and that exactly what these entities are is considered to be what differentiates the varying definitions (Saracevic, 2007). As mentioned above, in IR relevance is generally thought to occur between a resource or piece of information on the one hand and a query, topic, user or task on the other – or rather the information needs associated with these. In ordinary discourse, on the other hand, relevance occurs between a statement (or an idea) and another statement (or topic) or a situation or person. It is thus possible to equate the anchor points of relevance in the two cases, as in fig. 6.5. This parallelism is important when developing a relevance theory for IR, because it allows us to draw conclusions across the divide between IR and discourse.

IR	Discourse
User	Listener
User task	Listener situation
Query	Speaker A's utterance
Query topic	Utterance topic
Resource	Speaker B's utterance
Information in resource	Meaning of utterance

Fig. 6.5. Comparing IR and discourse relevance. Speaker A may for instance have asked a question or given a comment and subsequently needs to evaluate the relevance of B's reply.

According to Relevance Theory, a stimulus is relevant to a person if it benefits him or her by producing a *positive cognitive effect*, provided that the processing required in order to achieve that effect is not too great for it to be worth the effort. No direct reference is made to prior discourse. Thus, there is no strict requirement that a statement be related to the topic under discussion, and indeed people are free to make any statement they like at any time. The fact remains, however, that coherent discourse is the normal state of affairs (Thomsen, 1992; Eggins, 2004), and this is accounted for if we allow that processing will be considerably easier if the utterance is closely related to what has just been said. Thus, utterances are relevant, not to other utterances, but to hearers who have heard those utterances.

Transferred to search, this means that anything which the user might find useful or interesting would be a valid result. Results are relevant to the user and not to the query, and the query is merely a cue which gives a hint as to what kind of responses the user would currently find most interesting and easy to process. The task of a search engine is then in principle to provide results which are relevant to the user, in order of diminishing processing ease proceeding from the cue.

We can now truly begin to explain why the HRC applies: Stimuli which are at the forefront as matters under current scrutiny require less processing effort to retrieve and hence automatically qualify for higher relevance. Relevance to background information and long-term interests is costly to process and evaluate, while short-term interests are by definition within the current focus of attention and hence cheap. The task is in-between in accessibility, since it is presumably relatively recent, but not as recent as the actual request which expresses a short-time interest. This explains how the query can on the one hand be a mere cue and not a demand for a particular response while at the same time information that matches the query closely is under normal circumstances considered more relevant than something which is merely interesting in general.

Similarly, discourse is not coherent if people contribute different pieces of information which are unrelated but all in their own way relevant and productive of positive cognitive effects. Relevance Theory accounts for this by assuming that a stimulus is only optimally relevant if it is easy to process, i.e. relevance grows with the positive cognitive effect produced and diminishes with the effort required to process the stimulus. Obviously, what is currently under discussion is more easily accessible and hence cheaper in processing costs than something unrelated. Accordingly, contributions related to the immediately preceding discourse are not only more logical but also strictly more relevant. Perhaps this effect is at its strongest with questions; answering a question is so much more natural than contributing some other interesting information in that situation that we cannot but assume some very strong reasons for this behaviour and thus infer some meaning from this deviation from the expected behaviour which in the end will usually turn out to be related to the question after all (Leech, 1993). Whether a query is a question or not, queries share this strong presumption of an answer being the most natural reaction.

Wilson and Sperber (2004) focus on informativity and topicality in their concept of relevance as a basic principle of human cognition; in their eyes, relevance can be measured as the proportion of cognitive gains *vs.* processing efforts required, and all other measures of degrees of relevance are only proxies for this. This single-minded view of relevance is caused by a strong focus on cognitive effects. Whereas the present theory operates with strength, amount and ease gradations along all six dimensions of relevance, Wilson and Sperber effectively only recognize strength and amount of relevance in general on the one hand, together considered to represent gains, and cost of relevance in general on the other, and in both cases with a strong focus on informativity and topicality.

It would seem to be true that speakers mainly care about the coefficient of gains and costs. However, it is suggested here that neither is a simple variable. There are many tasks that require processing in a search session – scanning the results list, reading documents, assessing relevance continually at the various stages and in all dimensions, determining utility in the situation, and understanding the task in the first place. Gains are even more complex. A resource may provide cognitive as well as practical and aesthetic or entertainment gains when applied; even social gains through attending to certain people. Gains are increased when a larger amount of information is read, when more store is set by it because the author is trusted, when the reader is more attentive due to interest and desirability, and when the information is more directly topical and useful. The output is simple gains vs. costs, but the underlying functionality involves a complex set of interlocked cogwheels. Evolution may primarily have been attentive to the output, but if we are to reconstruct relevance in IR, then we need to take the machinery apart and count the cogs.

When the final binary choice of relevance is made, the various relevance values are compared to subjectively defined thresholds. Some are static; perhaps the user needs at least two pages of material, or only has time to read at most three pages, perhaps there is a limit to how much effort a given human can spend on the task, and so on. Other limits may be non-static, such as comparing the cumulative gains to the cumulative costs. This results in the effects noted by Wilson and Sperber and taken to be the central mechanism in relevance, while in fact it would seem to be a superficial effect.

Given that Relevance Theory has its origins in discourse, it is perhaps natural for its adherents to overlook the distinction between the various dimensions of relevance as some dimensions are by far more important than others in discourse relevance; a speaker need not continue talking on the same topic as long as his or her contribution is informative in the eyes of the addressee, and utility is not commonly considered to be necessary for an utterance to be relevant – or even a relevant parameter. An utterance certainly does not necessarily have to be elegantly expressed in order to achieve relevance. It does of course have to be spoken in a language understood by the addressee (i.e., be practical), but this is so much taken for granted that it is not included in the common notion of relevance for discourse either.

The present investigations into the nature of relevance may inform the understanding of the concept in discourse studies in several ways. Utility may indeed make an utterance relevant even if it is not informative on the topic on which the addressee expected to receive information; this is a well-known fact from pragmatics (Grice, 1989; Leech, 1993). Furthermore, desirability in the form of aesthetic expression as well as interestingness of content may well motivate an addressee to attend to a given utterance in preference to another one spoken by someone else at the same time.

> If a user values a witty style as much as an informative content, who are we to say that he is wrong? (Cooper, 1973)

Hence, as soon as there is a comparative aspect where the listener has to choose the *most* relevant of several available inputs, the situation is the same in discourse as in IR. Intentionality incorporates social relationships of power and affect between participants; we tend to listen more to our superiors and those who are dear to us, and this may be considered part of what makes an utterance relevant to attend to. Again, this becomes especially clear in a competitive context where we have to choose whom to listen to among several speakers.

Cognitive gains in the sense of Wilson and Sperber are a mixture of informativity and utility. Strictly speaking, only informativity is purely cognitive, while utility may refer either to cognitive or extra-cognitive advantages. The reason that Relevance Theory does not recognize this division of labour is that receiving a message is of course invariably a cognitive process involving the interpretation of information.

IR researchers usually talk of information needs rather than potential gains. This is because there is assumed to be some kind of need leading to the search in the first hand, even if it is not very clearly defined. In discourse, on the other hand, utterances are not supposed to satisfy the listener's needs (except in the case of questions). Rather, they provide information which can be added to the user's stock of knowledge. Even though an utterance does not satisfy a need, it cannot be on just any topic; it must be relevant, or in other words, intentional, practical, informative, topical, useful and desirable, at least to a certain extent. In a string of utterances, each utterance is the clue for the relevance of the next one in that each defines a topic which is taken to be the default for the next. Maybe this is the way it ought to be understood in IR as well. Needs should probably be replaced by potential information gains. This would explain why the satisfaction of an information need can only be ascertained after the search is completed; only then has the information been gained. The query is merely a clue to guide the LAT to a relevant response. The system may in principle respond in any way, just like a speech act participant, but it should avoid giving irrelevant responses, and this is achieved through attending to the query.

The methodological considerations presented in section 1.1.2 above apply equally to studies of language in general. In a discourse exchange, it is possible to investigate either how the relevance of person B's reply varies with person A's initial utterance, in what context B's reply would be relevant to a given utterance by A, or what expression speaker A might use in order to receive a desired kind of response. Thus, we may once again define the parameters *Q* (the expression uttered by speaker A), *C* (the context relevant to *Q*), and *D* (the desired response to *Q*). As with ideal resources in search, the desired response is not an actual utterance, but an input which will, in Wilson and Sperber's words, produce a positive cognitive effect. The three possible types of discourse studies corresponding to those proposed for IR are the following.

- **Relevance and rhetoric studies**: It has been common – especially since Grice (1975) – to study what responses *D* are relevant in a given discourse context (which includes both *C* and *Q*). In this setup, the initial expression uttered by A is simply considered a given part of context, while focus is on B's reply.
- **Context and expectation studies**: Alternatively, it is possible to explore in what context *C* speaker A uttering *Q* would be interested in, or expect, the response *D*. This setup does not explicitly study the fulfilment of A's expectations (which is not to say that linguists that do this kind of research ignore this matter; they simply combine the context and expectations studies with what we have called relevance and rhetoric studies, and do so successfully).
- **Utterance expression studies**: Studying actual linguistic expression means examining what expression *Q* a speaker in context *C* would choose given a wish or expectation of receiving the response *D*.

We may now go on to define the three relations in discourse corresponding to the *need*, *request* and *search* in information retrieval.

The **need** is a relation *n*(*D*,*C*) between a context *C* and the response *D* that is relevant to the speaker A(*C*) in the context *C*. The relation is such that *D* comprises exactly those responses needed by the speaker in context *C*. Note that unlike information retrieval where many resources are normally returned in a list format, there can normally only be one response to an utterance in discourse, unless there are several respondents.

The **utterance** (corresponding to the *request*) is a relation *r*(*C*,*Q*) between a context *C* and an expression *Q* uttered by speaker A(*C*) in the context *C*. The relation is such that *C* comprises exactly those aspects of context which are relevant for the formulation of *Q*.

Finally, the **move** is a relation *s*(*Q*,*D*) between an expression *Q* and the response *D* that is relevant to speaker A(*C*) uttering *Q* in the context *C*. The relation is such that *Q* reflects the need for the response *D*.

The context *C* dictates what responses *D* are needed or expected, the expression *Q* defines what aspects of context are relevant to it, and the desired responses *D* are reflected in the expression *Q*.

6.5.2 Consequences for tool design

Based on the work of De Mey (1980), Ingwersen (1992, p. 22) suggests that there are four stages in the evolution of IR systems: "1) one book[138] = one assigned class or index term, or single term extraction from the text, 2) keyword phrases, morpho-syntactic term extraction, clustering, 3) semantic values combined with request modelling, 4) really adaptive, knowledge-based systems, pragmatic systems". The fact that these seem to fit the four central facets of queries (substantial, formal, representational and interpersonal) may explain why there should be four stages. It is of course possible alternatively to classify systems according to their ability to include the various facets of resources. For the four central facets, the result would be similar to the above, though notably there may be a huge divergence between a system's ability to interpret queries and resources. However, there are (at least) two more facets, so that in fact there are six potential stages of sophistication. The first is hardly of any interest, representing as it does merely a system able to provide access to a previously defined resource with no actual search, but the sixth facet involves systems that are able to include affective characteristics and prioritize resources that a particular user is most likely to find pleasing either as regards its looks or its information contents, etc. In principle, it represents the ultimate information retrieval system which would always return what the user desired most. However, as that goal is obviously not easily achievable, aiming for it may not be wise, or at least should be done with due caution. After all, most users would prefer to be given what they ask for rather than what the system *wrongly* presumed would be of most use and interest to them.

It is in general tacitly assumed by the IR community that a search engine should ideally be able to respond immediately to a request and should aim to deliver the correct result after the first query has been submitted. However, research into relevance dynamics shows that this may not be possible or indeed desirable, since several searches are often *required* in order for the user to even understand his

[138] I.e. one resource; the example originates in a library context.

or her task and needs (Vakkari & Hakala, 2000). As pointed out by Swanson (1977), search is essentially a trial-and-error process, or, as Taylor (1962) put it, "when the system has answered a question, the process has just begun for the inquirer. His need for information is not at an end; it has only changed"[139]. Thus, rather than delivering the correct answer immediately, IR system designers should focus on helping the user understand the task and define the information that is needed.

A search engine should help the user understand the task and define the information that is needed.

Cooper (1971, p. 29) pointed out that, in contradistinction to an inferential system, with a search engine "the user and the system must work as a team to supply the answer to the question". He meant this to imply that such a system is unable to answer the question all by itself but instead must enrol the user and exploit his or her cognitive abilities by providing relevant material. Today, we might need reminding that the opposite holds: We now tend to think of search engines as purely information-providing systems, when in fact they may help actively in solving the task by suggesting possible ways of moving on and getting to the next stage of the process. IR systems could be so much more than just information retrieval systems; they could be task-solving partners.

Navarro-Prieto *et al.* (1999) discuss the *computational offloading* provided by a system interface, i.e. to what extent the system and its interface reduce or increase the amount of cognitive effort required of the user in order to fulfil his or her task. A well-designed system reduces the efforts on the part of users by doing some of the work for them. This could apply to more than layout and user-friendliness if the search engine took an active part in the user's project of achieving a goal. There is, however, a snag to this user-friendly approach: Users have been shown to prefer quick and straightforward searches, even when they would in fact benefit from a more sophisticated system (Azzopardi, 2009). It is clear that users should always be given the fastest response possible, but at the same time they might also be offered a chance to refine the search by more advanced methods such as coaching dialogues ("wizards") or just the option to specify particular needs and disambiguate terms.

6.5.3 Context and facets

In light of the Faceted Stimulus model and the Regulated Flux model, the idea of collecting contextual information may have to be rethought; each facet may be relevant to the user, and so information on

[139] Turpin and Hersch (2001) found that users on average issue about 3-4 queries in a session, depending on the difficulty of the task and the sophistication of the system. Smith and Kantor (2008) suggest that users are able to compensate for poor systems by issuing more queries.

each facet may need to be collected. The intentional resource may be relevant because the user belongs to the group of readers for whom it was created, or because he or she was told to read the document; the substantial resource may be relevant because of its physical characteristics; the formal resource may be relevant because of providing information or value; the representational resource may be relevant because of contributing information on a certain topic; the interpersonal resource may be useful because of providing knowledge or value that is useful in the situation that the user is in; and the effective resource may be relevant because of being what the user currently needs. For each of these assessments to be anticipated by the system, information must be collected, and most of it is contextual (depending on the contextual core chosen).

It may not be practicable to collect all the contextual information that we would like, but we should at least have an idea of the kind of information to which we would ideally like to have access. This is the purpose of the present discussion; we shall attempt to develop an understanding of context and a typology of potentially relevant information that we might consider collecting in certain circumstances. Each relevance dimension begs the collection of a certain kind of contextual information from the user as well as from the resource. It is up to the designer of the system to decide what information to focus on, depending on the users' general needs and preferences.

- **Desirability**

  If you want a system to be able to anticipate desirability, you will have to collect information on the user's preferences as regards resource characteristics. These can then be stored in a long-term user profile, possibly one that the user is allowed to see and edit. This is especially useful in intranet search where users can stay logged in more or less permanently. Having got this information, it is of course also important to collect information on the features of resources, and this may not always be straightforward, because preferences may concern writing styles and psychological reactions to certain topics or images, etc. Satisfying such criteria will be very challenging.

- **Utility**

  If the system should be able to assess utility, contextual information on the task situation is required. This is something that will have to be collected explicitly from the user through some kind of interface which allows the user to choose from among a set of situations related to the request, perhaps ideally as a post-filtering opportunity for refinement. Once again, collecting information on the kind of situation in which the text is useful is likewise necessary; otherwise the user information is useless. This is not easy to derive automatically from texts.

Consequently, such a solution would benefit from collecting intentionality information directly from authors (see below).

- **Topicality**

Topicality is famously at the centre of current LAT ability and does not necessarily require the collection of any supplementary information from the user. However, it does necessitate semantic annotation of resources in order to be ideal. Semantic search with synonyms and related terms is also recommendable in order to allow for topic domains rather than mere labels.

- **Informativity**

Collecting information on user abilities is often problematic, but at least language preferences can be stored in a personal profile. In some cases, a certain level of expertise or understanding of certain kinds of resources may also be registered, but then texts must equally be tagged with corresponding levels, either through automatic analysis or manual registration.

- **Practicality**

Practicality information such as accessible formats and acceptable cost levels can be dealt with relatively easily if this is deemed necessary. Another kind of practicality information is whether the user in question is allowed to access a resource.

- **Intentionality**

If you want the system to know whether the intentional resource is relevant, you will have to collect information on the type of reader that the author was envisaging, and also the kind of situation in which the resource was thought to be of potential use to the reader. In addition, it would be necessary to know what kind of situation the reader is in, in order to compare this to the author's intentions. This is in fact a highly useful kind of contextual information.

It is generally – but wrongly – assumed that users are simply interested in any information that will satisfy their information needs. The problem stems from the very notion of *information need*, because it suggests that people merely need information when in fact they will only be benefitted by the *application* of information to their task. As noted by Bruner (1986, p. 53), an idea is not understood "by means of a positivist archaeology in which everything particular about it and everything leading up to it are finally dug up, labelled and collated. However much we dig and delve, there is still an interpretive task". Users need to perform the application themselves, so the information must not only be informative but also easily applicable. It is true that we often content ourselves with any information available because we have grown accustomed to the fact that this is often what we can expect.

However, it would seem that what we ideally want is information intended to inform us on the topic that we are searching for or to help us in the situation that we are in, resources which merely happen to provide the required information being second class resources that we do not trust as much.

For an example, imagine a user wishing to locate a nice café in Copenhagen and therefore issuing the query *cafés in Copenhagen*. The search might be imagined to provide the following three results, among others.

- A site by the name of *Dining out in Copenhagen* containing the sentence *We recommend the following cafés in Copenhagen…*
- A social site with the sentence *I have discovered a wonderful little café in Copenhagen called...*
- An extract of a novel containing the sentence *He grabbed his coat and ran down the street, hurrying to visit his favourite cafés in Copenhagen…*

In most contexts, the user would be much more likely to be satisfied with the first of these resources than with the others, with the last one being decidedly unhelpful. The second might be useful, but the first would be preferred, because it gives the overview and was intended to help people in exactly the type of situation in which the user finds himself.

Now imagine a different user who is experiencing pains in his chest and consequently chooses to consult a LAT and submit the query *chest pains*. The following responses might result.

- A health information site with the sentence *Pains in the chest may be due to the following factors…*
- A social site with the words *Lately, I have experienced severe chest pains*.
- A fictional text with the sentence *The deep wound in his chest caused him severe pains*.

Again, most people would probably prefer the first of these resources since it was written with the explicit intention of informing readers about real-life chest pains. We must conclude that information on the user's intention is not sufficient; it needs to match that of the informer. The LAT needs to know which documents represent such an intention. Obviously, the informer cannot be asked at the time of search, but the user can.

6.5.4 Conclusion

This second part has discussed some of the many problems relating to understanding the controversial phenomenon known as *relevance*. The Regular Flux model has been presented as a feasible model which brings together all the different manifestations of relevance and places them in relation to each other. It consists of six separate but interacting dimensions representing different ways of evaluating a resource. In Part III we shall find that there may be occasion to introduce another two dimensions of an abstract kind representing the type of stimulus assessed and the type of relevant contribution offered by it and corresponding to two additional facets; this calls for further research. Each dimension may be judged against three different scales which also interact in interesting ways. Furthermore, relevance in each dimension may be evaluated according to any of the three Schützian spheres of attention, and each of the six facets of a resource may be judged separately. In practice, people are only aware of performing a few of these assessments, and many of the tests involved in the model are simply *irrelevant* in any given case, thus relieving the user of some of the mental work. Even so, relevance assessment is a demanding task. It would therefore be a huge asset if IR systems were able to help with the process, but unfortunately this is far from straightforward, partly because of the immense complexity of the phenomenon and partly because it hinges on subjective assessment which is inaccessible to the system. Even if users were questioned directly, it is likely that they would in general be unable to answer, because so much of relevance assessment happens subconsciously.

This part has provided a framework which allows discussions to be had and tests to be made, and hopefully has at least contributed to an improved understanding of this fascinating phenomenon. In future, separate studies should be made of each relevance dimension. Not least desirability, utility, topicality and informativity merit thorough investigation in order to chart the effects of certain features of resources and of context that affect relevance judgments. It might be possible to test the processual part of the model by measuring assessment times; in principle it should take longer time to process many dimensions than few, and so the HRD and the principles of regulated flux outlined in section 6.2 might perhaps be the basis of such an experiment. Part III deals specifically with users' expression of their needs through queries.

Part III

The meaning of a request

Chapters 7-10

7 Identifying a need

> *What we need, at least in part, is a framework for investigating document retrieval systems – a framework which identifies the major types of possible retrieval situations. Without such a framework, the document retrieval problem is simply too large and the factors influencing good design are too varied to permit us to investigate these problems systematically.*
>
> (Blair, 2002a, p. 283)

There are many different kinds of search; some users look for resources that they have already seen, others for something that is expected or hoped to exist, still others go on a hunt for anything that will satisfy their needs. If a system is to return useful results, it must know (or try to guess) what the user intended, so various possible intentions related to a single query form must be charted and described. This is the challenge that is taken up in this final part of the thesis.

The IR literature traditionally places great emphasis on **user needs** as representing what a search system should strive to satisfy. Exactly how the notion is to be understood is rarely made clear, but it is generally thought of as "the perceived need for information that leads to someone using an information retrieval system in the first place" (Schneiderman, Byrd, & Croft, 1998). This would intuitively seem to be a reasonable definition, except that not all searches are triggered by a need for *information* in the narrow sense. However, the exact character of a need is not at all clear; is it an emotion, a state of lack, or something that you require? Or is it in fact all of these and more? The answer probably depends on what you require the notion for.

It would of course be wonderful if we were able to satisfy that ultimate need behind a search, but unless some kind of dialogue is instigated, we can only hope to know what the need is to the extent that it is expressed in a query, or at least follows logically from it when interpreted in context. Consequently, what we are interested in here is the way a need is expressed, whether in the form of a reference to emotions, problematic situations or states of lack, or possible solutions and desired resources, and hence our understanding of needs is one that takes query expression as a point of departure.

It is a well-known fact that query log analysis can at best reveal user intentions partially; firstly, because users may fail to communicate their needs appropriately; secondly, because they may choose to go for a partial solution to their perceived problem only (Teevan, Alvarado, Ackerman, & Karger, 2004); and thirdly, because any given query may be interpreted as representing various needs due to linguistic ambiguity. Thus, "while work has been done to connect loggable behavior with task classification, the success rate for such models, even for very general task types, is quite low" (Grimes, Tang, & Russell, 2007). Progress in this direction can only be achieved by settling on a different typology of tasks that is liable to be reflected to any extent in actual query expressions. This means uncovering a set of distinctions that represent the choices made by users when deciding how to phrase their queries.

As a provider of search engine technology, you want to help the end users of those systems as much as possible, whether they intend to buy products from a web shop, find information on some topic that interests them or decide to look for job postings. However, assisting people is rarely straightforward, due to the ambiguity of the queries they issue. Most queries are ambiguous in one way or another; some have ambiguous semantics and others are simply issued for entirely dissimilar purposes by users in different situations. Here is a typical example.

7.1	**sas**	WID (87)
	SAS	WID (44)

In a job search portal such as Workindenmark, this query could refer to a multitude of abbreviations such as the Scandinavian airline (as a place to work) or the statistics software (considered as a skill). Even if we assume that it refers to the airline, in a broader search context it may represent either a wish of going to their homepage, reading about the company, finding out what has been said about them in the media recently, getting some advice regarding travelling with them, seeing a list of ticket prices, understanding what the abbreviation stands for, checking that the name is correctly spelled, finding out how commonly it is mentioned in some site or in general, or doing something entirely different. There is thus no *simple* relationship between the form of a query and the intentions behind it. Presumably there is some relationship, although certainly not an unambiguous one, and knowledge of the relevant domain certainly helps. This part of the thesis explores that relationship.

With the unique Ankiro One Line Search technology, the Ankiro job search systems are able to identify various classes of information in a query, such as locations, working hours, positions, companies and terms referring to working environments, etc. This enables the system to suggest jobs that match the requested characteristics, but also to suggest related jobs and positions in nearby locations or with similar conditions, even if they are not identical to what was asked for. This is a very powerful tool and it has proven valuable to the end users of those systems. Nevertheless, the system can obviously not distinguish between intentions not expressed in the query, and the question is how queries such as the following are really to be interpreted.

7.2	**engineer copenhagen**	WID (24)
7.3	**student copenhagen**	WID (18)

There is nothing in these queries to suggest whether the users who issued them were engineers and students from Copenhagen looking for suitable jobs in the vicinity, or whether they were just people looking for positions as engineers and students in Copenhagen, with no information given about their current whereabouts or educational background; it is simply not possible to tell from the form of the queries which sense was intended[140]. Another example is problem-oriented search in a support environment, such as the one provided by the TDC site. Customers regularly search for help with various difficulties they have encountered. Again, it is sometimes impossible to know whether the query refers to a problem or its solution.

7.4	**buzz off**	TDC (1)

This could mean either that the user wants to turn the buzz off, or alternatively it might be a description of the problem that needs solving, in which case the user in fact wants to turn it *on*.

The following examples have an identical structure consisting of an adjective referring to speed followed by a noun, yet they represent entirely distinct intentions in this regard.

7.5	**langsom pc**	'slow pc'	TDC (259)
7.6	**hurtig pc**	'fast pc'	TDC (48)

[140] It is even conceivable that the two terms in a query may differ in the choice, so that one describes the current situation and the other the desired job, but this is probably less likely than an agreement across the terms within one query, and no queries have been interpreted that way below, unless specific information to this effect is present.

While the former describes a problem that needs solving, the latter is more likely a request for information on a new computer which will solve that very problem. In some cases, the decision seems easy enough to make when a query log is analysed, but that is only because we as humans immediately understand that the situation referred to is either problematic or ideal and as such is most likely to represent a problem or its solution; teaching a computer to tell the difference is going to be far more difficult. The following query, for instance, clearly represents the problematic situation rather than its solution, but this is not apparent to the system.

7.7	**mobil i stykker**	'mobile (phone) broken'	TDC (7)

Of course, this is not likely to be problematic unless there are resources with specific instructions on how to take the appliance apart which could be mistaken for relevant information in this connection. However, the job search scenario shows that it can be crucial in some contexts. At the very least, we need a theory and a framework for discussing such cases.

One way of achieving a classification of queries based on user intentions has been to group them on the basis of characteristics such as the number of words they contain and the presence or absence of certain suggestive words or concepts on the basis of an assumption that these reflect differences in intention. This approach has met with some success, but only within certain rather narrow limits, as we shall be seeing shortly. It obviously necessitates a theory that defines the kinds of intentions available to choose from in the first place, and no satisfactory framework has as yet emerged, in spite of the endeavours. This part of the thesis is an attempt at filling this void by suggesting just such a model.

7.1 Query classification

Current IR tools only cater for a fraction of the diverse uses to which they are put, and it is vital that more uses are included in the design of future systems (Blair, 2002a; Allan, Aslam, Belkin, & al., 2003). However, even defining what an intention is, and deciding where to draw the line between *different intentions* is highly controversial. Various researchers have responded to this challenge by undertaking a classification of queries into types reflecting user goals, but unfortunately it has proven just as difficult to decide on a typology of goal types as of intentions. Even defining what a goal is proves less than straightforward. Before suggesting our own new model in chapter 8, we shall briefly examine the current state of the art in query classification along with some alternatives suggested in the recent literature.

7.1.1 The state of the art

Coming mostly from computer science, information retrieval researchers in this subfield do not normally perform any in-depth linguistic or cognitive analysis of queries, which are generally treated as little more than strings of words to be matched[141]. This is certainly a considerable improvement from treating them as strings of characters, but there is so much more to a query that remains unexplored and unexploited. To be sure, in one type of studies which has achieved wide and often perhaps too unquestioning acceptance, a very shallow analysis is undertaken in order to determine whether the user intended the query to be *transactional*, *navigational*, or *informational*, as suggested in a rather fragmentary but nonetheless seminal and highly influential paper by Andrej Broder (2002)[142]. To this end, the number of words in a query may be counted, and the presence of question words as well as of certain lexical items that signal a transactional intention (*buy*, *download*, *software*, etc.) is noted (Tamine-Lechani, Boughanem, & Daoud, 2010). Unfortunately, the enterprise is hampered by a failure to distinguish between what we shall call a user's **Aim** to generate a certain response from the system and the **Goal** that he or she wishes to achieve in the end on the external (work task) level, as well as other distinctions which are completely overlooked. This has led to incompatible definitions of the three types, resulting in some confusion and corresponding incompatibility of study results. Also, the entire question of what user intentions represent in the first place is largely left unanswered. The following is an attempt at remedying these deficiencies. In the end, we shall find that a different approach is called for where the strategy is to make no attempt at capturing user intentions beyond what is directly recoverable from query form. Rather than trying to classify intention types, we shall be satisfied with typologizing over types of search behaviour which signal intention. We shall argue that this is the only viable approach at present unless real-time user feedback is available, which is mostly not the case.

In a survey of AltaVista users, Broder (2002) found that roughly 25% of searches were reported to be navigational, while an estimated 36% were transactional. Remaining queries (39%) were assumed informational. In a comparative log analysis, he found that 20% of queries were navigational and 30% transactional, with the remaining 48% assumed to be informational (*sic*; presumably the last 2 % are the result of rounding errors). These numbers are slightly misleading in that while question 2 in his survey identifies navigational *Aims* (as well as non-navigational Aims which are part of an informational

[141] The little note that *is* made of linguistic features in query classification generally appears to hark back to Mariam Daoud's doctoral dissertation (Daoud M. , 2009) and is summarized in Tamine-Lechani *et al.* (2010).
[142] I refrain from giving definitions of the three query types here, for reasons that will become clear in the following; authors simply use a range of divergent definitions without realizing it, and so no standard definition can be given, unless Broder's descriptions are to be considered the standard by definition.

Goal), question 3 in fact singles out transactional *Goals* (fig. 7.1). In other words, they do not concern the same phenomenon.

2. Which of the following describes best what you are trying to do?
    a. I want to get a specific website that I already have in mind
    b. I want a good site on this topic, but I don't have a specific site in mind
3. Which of the following best describes why you conducted this search?
    a. I am shopping for something to buy on the Internet
    b. I am shopping for something to buy elsewhere than on the Internet
    c. I want to download a file (e.g., music, images, programs, etc.)
    d. None of these reasons

Fig. 7.1. An excerpt from the questionnaire used by Broder (2002).

Note especially that this survey effectively causes queries with a navigational Aim and a transactional Goal (i.e. buying or downloading from a predetermined site, etc.) to be classified as both navigational and transactional at the same time. Nevertheless, the results are interesting and suggestive of the approximate proportion of navigational and transactional queries.

Jansen *et al.* (2008) report that more than 80% of the queries in their log study were informational, which is clearly at odds with Broder's results indicating a proportion of less than half. The numbers arrived at of course depend as much on the criteria employed as on the resource collection chosen for the test. Rose and Levinson (2004) found that, in their study, approximately 13% were navigational, 61% informational and 25% transactional. This is again significantly different from Broder's results, though not in the extreme degree of Jansen *et al.* The disagreement is likely to reflect the severe difficulties associated with determining user intent according to the existing partly inconsistent classifications.

It is not entirely clear whether the survey and the log analysis in Broder's paper were performed on the exact same data. Even so, it would seem that the log analysers were able to classify fewer of the queries into navigational or transactional than suggested by the survey. This is hardly surprising; it is likely that in the remaining queries the user's intention did not come through – or at least not clearly so – due to linguistic ambiguity (assuming that the analysis was sensibly performed). Setting aside the question of whether users are actually able to identify their intentions at all when participating in a survey, these results would seem to indicate that approximately 20% of navigational queries and 20% of transactional queries are not discernible from their linguistic form as belonging to the type in question.

However, if we can interpret the remaining 80% of queries correctly, we shall certainly have taken an important step in the right direction[143].

Typologies are always subject to the criteria chosen, and it is crucial to make the right choices here. Existing classifications of user intentions or goals are not even clear about their criteria, which makes it impossible to classify queries consistently according to the frameworks in question, as is evident from the highly divergent results arrived at in different studies. It is important that attempts at classifying queries according to assumed user goals and intentions make some reference to the actual query so that only properties which are expressed in the query in some way are used. If a query were not a linguistic message, it would not be possible to ask *why* the user made a certain request, yet that is exactly what Broderian query classification is about.

> If we imagine seeing the world from the perspective of a search engine, our only view of user behavior would be the stream of queries users produce. Search engine designers often adopt this perspective, studying these query streams and trying to optimize the engines based on such factors as the length of a typical query. Yet this same perspective has prevented us from looking beyond the query, as to *why* the users are performing their searches in the first place. (Dani, 2008, p. 8)

The reason that query classification should be possible at all is that being linguistic messages queries do indeed convey intentions, albeit frequently in an ambiguous and somewhat unclear manner. The challenge is to decode what hints there are to user intentions in the relatively short queries that people issue, and to avoid falling into the trap of interpreting queries that do not present sufficiently unambiguous clues as regards the intentions behind.

The three categories suggested by Broder (2002) can be informally described in a simple way, such as the following.

> [L]e type de besoin derrière la requête est défini comme étant informationnel lié à la recherche du contenu informationnel de documents, navigationnel lié à la recherche des sites d'accueil des personnes, organisations ou autres ou transactionnel lié à la recherche des services en ligne.

[143] Dani (2008) reviews attempts at operationalizing Broder's classification, something which we shall not do here.

> 'The type of need behind the request is defined as being informational when representing a search for the informational content of documents, navigational when representing a search for homepages belonging to individuals, organizations or others, or transactional when representing a search for online services.' (Daoud M. , 2009, p. 30; my translation)

In this quote, the difference between the types is presented as being one of *what you are searching for*, which makes it sound quite straightforward, but in fact when actual definitions are attempted it turns out that researchers understand the terms in slightly different ways, resulting in different classifications of actual queries. Some definitions are explored below.

- **Informational searching**

Informational searches may be defined with reference to either the kind of user need that trigger them or the kind of response that the user is hoping to achieve. The following two definitions differ in their choice here, but the ultimate idea is similar.

> Jansen *et al.* (2008): The intent of informational searching is to locate content concerning a particular topic in order to address an information need of the searcher. The content can be in a variety of forms, including data, text, documents, and multimedia. The need can be along a spectrum from very precise to very vague.

> Broder (2002): The purpose of such queries is to find information assumed to be available on the web in a static form. No further interaction is predicted, except reading. By static form we mean that the target document is not created in response to the user query. This distinction is somewhat blurred since the blending of results characteristic to the third generation search engines might lead to dynamic pages.

Jansen *et al.* focus on the information need of the user and point out that the actual response may not be crucial, or at least may vary considerably between searches. Broder, on the other hand, thinks that the kind of response resource is exactly what sets information searches apart – informational queries seek to retrieve information "assumed to be available on the web in a static form". To Broder, a request that involves the consultation of a database is by definition a transactional query, as the information is "created" on demand in response to information provided by the user. Thus, in a search for *weather in Copenhagen*, the user supplies the geographic parameter that enables the search engine to extract the required information from a database. Traditionally, such a search would have to steer the user to a site from which to consult the database. These days, the information can be retrieved directly by the search engine.

It is important in this connection to be aware of the fundamental difference between data and information. Static documents can be said to contain information (in one sense of the word), because they are intended to be read and interpreted, but databases do not store information because no one intended them to be read; rather, they are meant to be used for the independent construction of information. Accordingly, retrievals from a database are transactional and not informational.

The two definitions provided above are not incompatible insofar as Jansen *et al.* list static resources only. However, the authors do not care to include an explicit delimitation of possible resource types. The result is two entirely different definitions which cause highly different sets of queries to be assigned to these types, and lead to insurmountable confusion. We shall therefore have to separate the two definitions as delimiting types according to two distinct parameters; there can be one typology based on desired responses and one based on user needs[144], along with several others. Each typology contains a full range of types, some of which are more common than others. This should not lead us to excise the most frequently occurring types from various typologies and merge them into one general typology of "apples and oranges", however. Blending definitions based on incompatible criteria can only lead to meaningless classification and baffling overlaps between types in borderline cases.

● Transactional searching

Transactional queries can be defined on the basis of the user's ultimate goals or the type of response desired. Authors who use this kind of definitions tend not to be clear about the difference.

> Daoud (2009): […] l'intention est d'accéder à des services en ligne.
>
> 'The intention is to access online services.'

However, most researchers are inclined rather to focus on the actions that users are required to perform onsite in order to achieve their goals.

[144] Or another more precise notion such as the terms *Goal* and *Aim* introduced here.

Broder (2002): The purpose of such queries is to reach a site where further interaction will happen. This interaction constitutes the transaction defining these queries. The main categories for such queries are shopping, finding various web-mediated services, downloading various type of file (images, songs, etc), accessing certain data-bases (e.g. Yellow Pages type data), finding servers (e.g. for gaming) etc.

Jansen *et al.* (2008): The intent of transactional searching is to locate a Website with the goal to obtain some other product, which may require executing some Web service on that Website. Examples include purchase of a product, execution of an online application, or downloading multimedia.

A transactional query always implies some form of interaction in both directions; the user supplies some information and the system responds, either by retrieving information from a database or by causing a requested action to be performed. Such actions can be the delivery of non-informational wares, or the carrying out of a job.

Rose and Levinson (2004) split transactional queries (called by them *resource queries*) into four subtypes as follows.

Rose & Levinson (2004)

- **Obtain**: The resource is something I plan to use in the offline world, such as song lyrics, recipes, sewing patterns.
- **Download**: The resource is something that needs to be installed on my computer or other electronic device to be useful.
- **Entertain**: My goal is simply to experience (typically view or read) the resource for my enjoyment.
- **Interact**: The intended result of the search is a dynamic web service (such as a stock quote server or a map service) that requires further interaction to achieve the user's task.

This is quite reasonable, seeing as such searches are really very different in nature. The lack of consensus even made Lee *et al.* (2005) refrain from studying transactional queries at all, and indeed we shall not consider transactional queries to represent one uniform category either, as intentions vary so much as to render the category unavailing.

If the user wants to buy or download something, the job of a search engine is to lead the user to an appropriate page from which the transaction can be performed, or where necessary information can be exchanged before the actual transaction can take place; what the user needs in this case is guidance to a location. Truly transactional queries in the sense of a search that initiates a transaction directly have only become possible with recent technology; traditionally, most transactions are approached by way of a navigational or informational query. Even so, we shall see in connection with the empirical studies conducted here that it is in fact helpful to have a type that includes all queries where the user has transactional intentions, irrespective of whether additional steps need to be gone through en route. Indeed, it seems that for the purposes of query construction users treat the combination of search system and interaction site as together constituting “the system” and that being able to buy wares from a site reachable through search accordingly counts as being able to obtain them by searching. In other words, the surface expression of a transactional query is not significantly different from that of any other type, except in some particulars as regards the semantics of some words that may (but need not) occur in transactional queries, such as *buy*, *software*, and *jpeg*, etc. (see section 7.2.2), and this may of course be true of many other classes of query which have not been elevated to query type status. The present study focuses on query formulation, and what matters in this connection is what the user is aiming at through the query, and not what he or she actually needs to go through. Consequently, transactional queries shall not be considered to constitute a significant category for our purposes. Transactional searches may tentatively be defined as those which suggest a **complex system** where interaction with the search engine itself is not enough. However, even this definition would exclude cases where results are returned directly, such as information retrieved from a database and displayed among the search results, thus emphasizing the idea that transactional queries do not make up a homogenous category.

● **Navigational searching**

Navigational search intuitively constitutes a rather clear-cut category in that it consists of cases where the query simply refers to the name of a website that a user wants to be taken to. Even so, the following definitions do not seem to delimit navigational queries clearly from the above types as they do not make explicit reference to the form of the query but simply base the definition on what the user has in mind[145].

> Jansen *et al.* (2008): The intent of navigational searching is to locate a particular Website. The Website can be that of a person or organization. It can be a particular Web page, site or a hub site. The searcher may have a particular Website in mind, or the searcher may just 'think' a particular Website exists.

> Rose & Levinson (2004): We define the **navigational goal** as demonstrating a desire by the user to be taken to the home page of the institution or organization in question. To be considered navigational, the query must have a single authoritative web site that the user already has in mind.

Clearly, many types of search involve locating a website; what makes navigational queries stand out is that this desire – and not the wish to achieve something particular from the site after having got there – is what determines the linguistic form of the query.

With navigational queries, users simply want to be taken to a certain page which they know or assume to exist. The site is one from which they expect to be able to acquire needed information or where they intend to perform some action. Note that such searches do not refer directly to the desired information or action but rather to the site from which it is presumed to be available. The defining characteristic of a navigational search is thus that the user refers to the name of the website in question in the query, and that the entire point of issuing the query is to go to the site with no mention of what the user is planning to do once he or she gets there.

Rose and Levinson's definition reflects the opinion of Broder (2002) that navigational queries usually target a single web location (or at least one type of location), and can accordingly be assumed to have no more than one "correct" match. Lee *et al.* (2005) even decided to focus on the criterion of whether the user wished to visit the single correct resource that matches the request or was willing to read several suggested resources as being the distinctive feature when classifying queries into navigational and informational ones. However, since the user does not necessarily know the page in advance and certainly does not usually know what

[145] It is not entirely clear what Rose and Levinson mean by "the query must have a single authoritative web site"; it seems merely to imply that the user is thinking of a certain site, but it could be that they actually intended to make a reference to the expression.

other pages might be available, this criterion must be weakened accordingly. There may well be several pages answering the user's navigational needs.

According to the criteria mentioned in the literature for navigational queries, the partial URLs that people enter into a search box as a shortcut in order to get to the page in question should be included in that category. The present author does not agree with this assignment, since partial URLs are not references to a page itself but meta-references to its location on the web. Thus, they represent a special behavioural phenomenon which does not fall under any of these types. Since they are not linguistic expressions, they have been excluded entirely from the empirical analyses performed in this thesis.

Even from this cursory presentation of definitions it is clear that the three Broderian query types are defined on the basis of highly divergent criteria[146]. Transactional queries are formally defined on the basis of the required user behaviour but informally described as involving certain goals; informational queries are defined with reference to either the kind of goal or the type of response looked for (static information), and navigational queries are defined with reference to the Aim of getting to a specific page[147]. The crux of the definitional problem appears to be the meaning of the word *intent*. Jansen *et al.* clearly assume that if the intent is "to locate a particular Website", then it cannot be anything apart from that, i.e. the user *only* intends the query to retrieve a site. This will be considerably clarified with the proper introduction of the terms *Goal* and *Aim* in chapter 8 below; the Aim of a navigational search is *GETTING TO A CERTAIN WEBSITE*, while the Goal is basically unspecified in such cases as the purpose of going there is not spelled out. With informational queries, on the other hand, depending on the definition, the Aim may still be to locate a website or to find an interesting paragraph, etc., but the Goal is to find information or facts, or the Aim is to retrieve topical information, with unspecified Goals involving the use of this information. The definition provided by Jansen *et al.* even appears to require the simultaneous application of both criteria as they refer to "locat[ing] content concerning a particular topic in order to address an information need of the searcher". In other words, the categories of informational and navigational queries do not always belong in the same typology, because researchers are not careful to make sure that they refer to the same type of phenomena. In section 7.2, we shall consider some alternative classifications suggested in the literature and see if some of them might possibly fare better than the standard trichotomy.

---

[146] This was noted in part by Lee *et al.* (2005)

[147] In theory, it is possible to unite these definitions as referring basically to the kind of response aimed at, viz. interaction, information and website. This would however, to my mind at least, be somewhat contrived.

7.1.2 Aims and methodology

In spite of its shortcomings, the Broderian trichotomy remains useful as a coarse-grained general framework for understanding user intentions. However, even supposing that we were able to identify query types correctly and in a consistent manner, the shallow distinction between transactional, informational and navigational queries would still only be a meagre start. If possible, identifying the differences between queries looking for advice, information, facts, etc. would be far more helpful for the simple reason that it would provide a much richer array of possible interpretations. The existing classifications have too few categories, their definitions are neither consistent nor clear, they overlap, and the informational category in particular is largely negatively defined and should be broken down into a range of very different types, as will be clear from the data presented below.

With these concerns in mind, this third part of the thesis looks at the search process in a cognitive and communicative light in an attempt to identify some ways in which a thorough cognitive analysis of the linguistic facts might potentially improve the understanding of why people phrase their queries the way they do. There is a huge, un-mined source of interpersonal information buried in this expressive variation. Even so, exploiting it in actual IR systems is by no means an easy task, and so the present work should largely be considered a theoretical presentation of some of the issues from among which future researchers might select some to focus on with a view to constructing more linguistically aware LATs.

Fuller interpretation of linguistic queries is going to require advanced language technology such as natural language parsers and ontologies, but as these resources are already in use in advanced semantic search systems[148], it is not in principle an unachievable goal; it merely requires great attention to linguistic detail, including the choice of form. Based on observations from the analysed query logs we shall proceed to suggest a new model of the formulation of queries in chapter 8. The hope is that we may be able in future to decode queries and unravel this cognitive and linguistic production process, arriving at the motivations behind. This would, theoretically at least, allow identification of the perceived need that gave rise to the observable behaviour. We shall find that this is a highly ambitious and probably even impossible project generally speaking due to the cybernetic, bisubjective nature of search behaviour discussed in Part II, but understanding the phenomena involved will at least increase the likelihood that some queries may be more properly interpreted, at first through manual analysis, and later in an increasingly automated environment. It is likely that we can get much further than we are today, but only by understanding the choices that people make when they formulate their queries.

[148] Such as Ankiro Enterprise Search.

Classifying queries in an unbiased way according to what intents they express is going to require two things: Firstly, a model of what queries may possibly refer to; and secondly, an understanding of how such reference is made and how it can be decoded during query classification. The main goal at present shall be to provide a framework for understanding and analysing queries. This is achieved through three secondary aims, all of which are closely related.

1) Developing a model which helps understand why users express their needs in certain ways.
2) Providing a theoretical discussion of essential linguistic issues involved when users express their needs, i.e. identifying types of referential, syntactic, morphological and lexical variation in query formulation, and attempting explanations of this behaviour.
3) Developing a method for analysing query logs that derives its data from the linguistic expression of queries and thus allows for unbiased classification.

The framework that we shall develop to account for user needs and desires is a **Stage-based Intent model**, meaning that it is constructed around the various stages in the psychological dimensions along which users move during a search as well as when planning it. In the light of previous research such as the model of psychological developments during search proposed by Kuhlthau (1993) and the models of need-to-query development discussed in Part I it is likely, for instance, that users first experience a motivation and then become increasingly aware of individual features of the problematical situation, followed by more conscious deliberations regarding what would be needed, and finally a more or less planned search for resources. On top of this model we shall introduce the notion of **target**, i.e. the element in the model that a user chooses (often subconsciously) to refer to when constructing the query expression. The user may elect, for example, to produce a query that describes the problematic situation, to search for information on a relevant topic, or to anticipate the actual wording of an imagined relevant resource. With this analytical approach, faithfulness to the linguistic expression is ensured, and hence the analysis is performed on the same material as the LAT has to go on, which provides for results that are far more meaningful than an analysis driven by subjective assessment of likely intentions behind queries.

One important project consists in devising a new way of analysing log files in a linguistically informed way that minimizes the risk of ascribing intentions to users when they are not explicitly coded in the expression and hence can only truly be identified by asking the user directly. The method developed here shall be called **Target Analysis**. It utilizes the Stage-based Intent model and provides a link between it and the domain of the queries analysed so that contextual clues can be included. This is important because context is often crucial to the interpretation of a query due to the short and ambiguous nature of the query strings themselves. Also, even though we shall concentrate on features coded in the message from user to system this does not mean that everything is explicitly expressed, so contextual information is essential. Existing methods of analysis employ categories which do not allow unbiased categorization because they are based on the concept of user intentions, and intentions are not necessarily coded in the expression. In order to solve this problem, we shall need to introduce the notion of **intent**, which is meant to specifically cover intentions coded in the expression. Sometimes, intention analysis (in the wider sense) is necessary, and focusing on intent will make this enterprise more sensible.

It is not possible to derive a model directly from the empirical data, as only a limited amount of information is recoverable from surface form. Sophisticated psychological experiments would therefore be needed in order to prove the model, and such proofs are beyond the scope of this project. Hopefully, future research into the matter will be sufficiently stimulated by the present work into constructing a model which at least seems to explain the observable facts in a satisfactory manner to take up the task of devising experiments based on it. The model has been defined in its entirety as a coherent framework which, when applied to the data, explains the observed variation and illustrates the choices made by users when expressing their needs in the form of queries. The methods developed as part of this project have already provided valuable behavioural data for the managers of analysed sites which have helped them rethink their systems in view of the increased understanding of user needs and inclinations.

7.1.3 Coverage and expected validity

In order to maximize the coverage of generalizations made in this thesis the empirical data is drawn from two search logs that differ on many parameters (cf. chapter 3). The TDC log contains mostly Danish-language material, and the results are expected to be valid for Danish and other similar languages such as most Germanic languages. Unfortunately, the results regarding the contrast between imperatives and other verb forms would not seem to be transferrable to English, since there is no distinct imperative form in English as against infinitives. Likewise, it is unclear whether it can be transferred to Romance and other languages that frequently employ the infinitive as a polite imperative (such as many Indo-Aryan languages, for instance), since this would blur the effects considerably.

The WID log is mostly in English, but here the effect of language is considerably less than that of domain context, which is highly specific in the Workindenmark case. The actual categories that queries are sorted into in the study of that log in chapter 10 are tailored to this context and must be renamed and reconceived in a different setting, but the basic model of referential targets introduced below is expected to hold across domains and to be useful whenever search logs are analysed. One simply has to accept that it is necessary to know the domain in question intimately and devise an analytical framework suitable for the search system that one is working on, based on the generalized model developed here; it is not possible to apply the same categories across the board. This confirms the idea expressed in the introduction to Part I that this kind of analysis is not really suitable to web search because in such a system there is no identifiable domain. Of course, if a web search engine were to be constructed in such a way as to be able to identify the domain in some cases, the Target Analysis introduced here might be applicable even in that context.

It is not known whether the tendency to express yourself in a certain way in a query is universal or specific to distinct virtual communities of users who employ search engines in similar ways. It must be considered highly likely that principles of query construction differ along a range of demographic parameters such as geographic origin, social class, age and gender, as well as IT expertise and experience with search, etc. (Spink, Ozmutlu, & Ozmutlu, 2002). It is certainly known that different people have different strategies when consulting a search engine (Fidel, 1984). What is claimed to be true of the population represented in the search logs used in this study should accordingly not be taken to necessarily hold universally.

7.1.4 Structure of Part III

The remainder of **chapter 7** explores various possible alternatives to Broderian query classification based on intentions and looks at whether there are different “types” of search. It is possible to typologize over search in different ways, but only a typology based on the expression is truly useful because other types are fundamentally irretrievable. Hence, a different approach is needed.

In view of this, **chapter 8** develops a new framework which can be called upon to place the observed variation in an enlightening context and describe the correlations in question. The Stage-based Intent model unites what is known about the cognitive, linguistic and behavioural sides of search into one model which furthermore fits into the Faceted Stimulus model of resources and perceived situations and the context model proposed in Part I, as well as the Regulated Flux model of relevance developed in Part II.

Chapter 9 investigates a range of variational issues in the expression of queries in the analysed search logs such as morphological, syntactic and lexical choices and the choice of what to refer to in the first place. The chapter is rounded off by presenting an empirical study of the use of imperatives in search as opposed to infinitives and nominalizations. The study illustrates the complexity of these issues and demonstrates how hard it is to disentangle the various effects and tendencies when performing a query analysis. Although no definite proof can be given at this point, the study also seems to indicate that there is a systematic correlation between the choice of morphological form and the user's basic intentions hidden beneath the tangle of counter-effects, an observation which supports the suggested Stage-based Intent model.

Finally, **chapter 10** looks at possible applications of the suggested framework and demonstrates how a query analysis informed by the Stage-based Intent model can be carried out in a way that takes the expression as a point of departure. Given what was said above, the procedure is not an easy one, but much useful information can be derived from a careful Target Analysis of a search log, if correctly performed. Apart from this, the chapter sums up the thesis and points out where further research is needed.

7.2 Towards an alternative classification

The IR field is in need of a suitable alternative to Broderian query classification which is both psychologically feasible and technically detectable by a LAT. We embark on this endeavour knowing full well that it may not be possible to satisfy more than one of these demands at a time. There are various criteria that one might possibly use for such a classification, and this chapter explores some of them.

7.2.1 A formal representation of queries

Before setting out to study actual query variation in search logs we shall take the opportunity to briefly introduce the innermost layer of the formal notation employed here. We shall see in chapter 9 that queries contain a surprisingly large amount of morphology and syntax. The inner layer of the formal representation signals morphological choices and to some extent also syntax. Not all details are represented, as only certain characteristics of morphosyntax are salient in search. Forms are marked for word class, especially as regards verbs ($_v$), nouns ($_n$) and prepositions ($_{prep}$), which are particularly important for our purposes. However, noun phrases are only marked on the head noun in order to signal the extent of phrases without the use of brackets, which are reserved for other purposes. Also, if the entire phrase has been borrowed from natural language nothing can be said about it which does not also apply to natural language, thereby ceasing to be of specific interest to search studies. It is the fact that a natural language noun phrase has been imported in the first place that is interesting and not its composition.

The definition of *noun phrase* that we shall be using differs somewhat from that ordinarily adopted in linguistics.

In search, a <u>noun phrase</u> consists of a noun with its preceding modifiers, where either is facultative: (A*) (N).

The reason for this decision is that query syntax differs from natural language usage in some respects: A noun preceded by an adjective is usually one noun phrase, but when a noun is followed by an adjective in a query this normally appears to represent two juxtaposed noun phrases. Compare the following examples from the TDC log.

7.8	**trådløst netværk** **trådløs netværk**[149] {trådløst netværk$_{n.sg}$}[150]	'wireless network'	TDC (648) TDC (384)
7.9	**netværk trådløs** {(netværk$_{n.sg}$) (trådløs$_{n.sg}$)}	'network wireless'	TDC (2)

While the former is clearly one noun phrase, the latter can only be interpreted in this way by accepting a scrambling rule which does not apply in natural language Danish or English. The only common exception to this interpretation of the sequence of a noun and an adjective is naturally postposed modifiers such as resultative ($_{v.res}$) perfective participles or adjectives.

7.10	**iphone låst** {iphone$_{n.sg}$ låst$_{v.res}$}	'iphone locked'	TDC (5)

The participle here has a verbal role and implies an elliptic auxiliary rather than a copula, resulting in a stative resultative construction describing a problematic state that needs resolving. The difference is

[149] This form differs from the former by not adhering to the rules of gender agreement in the standard language. The high frequency of this lack of agreement may be suggestive of a linguistic change under way.
[150] The curly brackets envelop one referential target, for which see chapter 8.

not always very clear, as it depends on the interpretation of the semantics involved; it is only if the participle or adjective describes what has happened that it is a resultative construction.

The resultative construction implies a strong focus on the problematic nature of the situation because it involves a retrospective aspect[151] where the speaker looks back at something that happened in the past but resulted in a state that characterises the present. There is nothing in the form of the expression *netværk trådløs* that reveals whether this is to be interpreted in a resultative way or not, but if it were, it would indicate that being wireless was a state that the network had entered into, possibly representing a problematic situation. This is simply an unlikely interpretation, given the semantics of the expression. The following example, on the other hand, is likely to be resultative rather than simply descriptive of the smartphone in question.

7.11	**iphone 3 slukket** {iphone $3_{n.sg}$ slukket$_{v.res}$}	'iphone turned off'	TDC (2)

Obviously, there is always going to be a certain degree of arbitrariness in the precise formulation of queries, and hence an inverted word order may in some cases simply be due to the user's adding an adjective as an afterthought. In such cases, there would really only be one noun phrase semantically, but we shall have to resign ourselves to the fact that this difference is undetectable, and hence we elect to always interpret such sequences as consisting of two noun phrases[152].

To the part-of-speech tag in the formal notation is added a morphological marker such as singular ($_{sg}$), plural ($_{pl}$) or proper ($_{prop}$) for nouns, and imperative ($_{imp}$), infinitive ($_{inf}$), participle ($_{part}$/$_{res}$), present ($_{pres}$), past ($_{past}$) or passive ($_{pass}$) for verbs. Verbal nouns are treated as nominal ($_{nom}$) forms of verbs, as the semantic relation to verbs appears to be what matters in search. In fact, we shall find that verbal nouns largely pattern with infinitives (cf. section 9.4). Nouns may furthermore carry one of the markers definite ($_{def}$) and indefinite ($_{indef}$). This notational system will be developed further below as more phenomena are introduced. The principles are laid out fully in Appendix B. The point of introducing a formal representation is to make explicit certain characteristics of individual queries in order to enable discussion of query variation and its causes. It is not claimed that the chosen notation is necessarily the ideal way of representing queries; it is merely meant as a modest start on which to build in future.

[151] Cf. section 8.1.4.
[152] Some afterthought additions may be identified through session analysis if the modifier was added after a first search was performed in order to make the term more specific. We have no such data for the material investigated here, for reasons laid out in section 3.1.

7.2.2 Query type recognition

The driving idea behind research in query classification is that it should be possible to recognize query types automatically by attending to features of the query, in combination with supplementary contextual information (Lee, Liu, & Cho, 2005). However, as noted by Broder (2002, p. 7), "inferring the user intent from the query is at best an inexact science, but usually a wild guess", and with the added complication that some queries refer to the user's Aims, some to the Goals and some to other stages of the search it would seem to be a considerable challenge. Clearly, if the definitions employed are based on such variable parameters as they seem to be, figures are bound to differ between studies, as indeed they do. Previous studies have identified some of the most important characteristics that contribute to the probability of a query belonging to a certain type. The following is a synthesis of this work based on Jansen *et al.* (2008), Daoud (2009) and Daoud *et al.* (2009), and representing the state of the art.

- **Transactional query characteristics**
  - **Length**
    - Query length (i.e., number of terms in query) > 2 (contains at least the transaction verb, the object and optionally an additional query term for explanation)
  - **Reference**
    - The query contains a reference to a transactional entity (e.g., *lyrics*, *software*, *pictures*, *games*, *recipes*, *movies*, etc.)
    - The query contains a file extension related to transactional entities (*jpeg*, *mp3*, etc.)
    - The query contains a transactional verb (e.g., *download*, *buy*, *reserve*, etc.)
  - **Language use**
    - The query contains a (transactional) verb
  - **Resource collection**
    - The query is specifically directed to images, sound or videos

- **Informational query characteristics**
  - **Length**
    - Query length (i.e., number of terms in a query) > 3
  - **Reference**
    - The query contains references to information storage (e.g., *list*, *playlist*, *text*, *paper*, etc.)
  - **Language use**
    - The query contains a question word (e.g., *how*, *where*, *what*, *ways to*, etc.)
    - The query uses natural language terms
  - **User behaviour**
    - The query is not the first submitted in a session
    - The user views multiple results pages

⬤ Navigational query characteristics

- **Length**
  - Query length (i.e., number of terms in query) ≤ 2
- **Reference**
  - The query contains the name of a company/business/organization/person
  - The query contains a domain suffix
- **Language use**
  - The query does not contain a verb
- **Resource collection**
  - The query is not specifically directed to images, sound or videos
- **User behaviour**
  - The user views the first results page

Significantly, Jansen *et al.* add that a query is considered informational if the criteria for the other types fail. There is thus no requirement that an informational query be necessarily longer than three words. There is also no guarantee that a given query belongs to one category only; quite on the contrary.

It is clear from the above that the criteria hitherto suggested are not primarily linguistic or psychological ones; much more could be done as regards interpreting the actual message conveyed by the query. Some reference is certainly made to language use, but only in a superficial way, and it is clear that much more in-depth analysis is needed. The linguistic analysis is limited to checking whether the strings contain question words, verbs and "natural language terms". The use of question words in queries is limited, and consequently that particular criterion can hardly be expected to contribute greatly to the number of correctly interpreted queries[153]. The use of verbs is also much more complex than the points above suggest, as we shall see below. The criteria above state that certain verbs are often present in transactional queries, whereas usually no verbs are employed in navigational queries. In light of this, one may be surprised to find that verbs appear quite frequently in what would be interpreted as informational queries concerning help on how to do something or have something done[154]. Here are a few examples from the analysed logs using a selection of inflectional forms.

7.12	**voting** {voting$_{v.nom}$}	WID (1)

[153] The TDC log contains 0.3% questions of various types (see section 9.3.2 for details).

[154] Daoud (2009, p. 120) acknowledges this but does not specify the role of verbs in informational queries. She simply states that verbs are "une propriété générale des requêtes informationelles ou transactionnelles, qui consiste souvent en un verbe représentant respectivement le besoin en information ou le service transactionnel demandé".

7.13	**working at age of 15** {working$_{v.nom}$ at$_{prep}$ age$_{n.sg}$ of$_{prep}$ 15$_{n.num}$}		WID (1)
7.14	**living in denmark** {living$_{v.nom}$ in$_{prep}$ denmark$_{n.prop}$}		WID (1)
7.15	**attach my cv** {attach$_{v.inf}$ my cv$_{n.sg}$}		WID (1)
7.16	**se tv** {se$_{v.inf}$ tv$_{n.sg}$}	'watch tv'	TDC (355)

These first examples clearly represent topical, informational queries, yet they just as evidently contain verbs. Even the distinction between nouns and verbs needs looking into, as in my experience nominalizations based on verbs also count as verbs in query construction (See section 9.4). This may not be very clear in English, and hence not as important as it is in Danish where nominal forms are much more distinct from the verbal forms due to a somewhat richer morphology[155].

Some queries containing verbs may be interpreted either as informational queries on a topic or as what might be called transactional queries in that they typically involve further interaction with the system.

7.17	**udskriv regning** {udskriv$_{v.imp}$ regning$_{n.sg}$}	'print bill'	TDC (1)

Referring as it does to the topic of printing a phone bill this query may be interpreted as being of the informational type. However, as it contains a verb it is also eligible as an instance of transactional search; the user may have intended to perform the action directly as a consequence of searching and going through a few intermediate steps. In fact, the latter seems a most likely interpretation in this particular instance, especially as it is expressed in the imperative, which intrinsically signals the wish to have an action performed. There are many other ways of expressing this particular need, not all of which contain a verb in the strict sense, although they do relate to the same linguistic root. Firstly, it can be stated in the form of a question as in the following example.

[155] In technical terms, Danish is much less prone to *zero derivation* than English.

7.18	**Hvordan udskriver jeg den fremsendte fakt.?** 'How do I print the forwarded inv.?' {Hvordan udskriver$_{v.pres}$ jeg den fremsendte fakt.$_{n.sg}$?}	TDC (3) [156]

There are also other variants containing verbs, such as the following.

7.19	**udskriv faktura** {udskriv$_{v.imp}$ faktura$_{n.sg}$}	'print invoice'	TDC (1)
7.20	**udskriv regninger** {udskriv$_{v.imp}$ regninger$_{n.pl}$}	'print bills'	TDC (1)
7.21	**print regning** {print$_{v.imp}$ regning$_{n.sg}$}	'print bill'	TDC (2)

These queries all contain imperatives; there happen to be no infinitives. This is likely to be a coincidence, though, as a broader study of the log across user intentions shows that infinitives usually do occur in contexts that allow nominalizations to be used, and nominalizations are plentiful in this case.

7.22	**udskrift af regning** **udskrift af telefonregning** **udskrift af faktura** {udskrift$_{v.nom}$ af$_{prep}$ regning$_{n.sg}$}	'printing of bill' 'printing of phone bill' 'printing of invoice'	TDC (4) TDC (3) TDC (1)
7.23	**regningsudskrift** {regningsudskrift$_{v.nom}$}	'bill print(ing)' [157]	TDC (1)

[156] The shown query is decidedly odd in that what is arguably the most important word – *faktura* 'invoice' – has been abbreviated. It is highly suspect that it has been submitted three times in this form, and this is likely to be a case of the user viewing multiple pages, cf. section 3.1.1.

[157] *Udskrift* is a nominalization derived from the verb *udskrive*. It can be interpreted in this context either as a nominalized action verb or as a noun describing the product of the action.

7.24	**genudskrivning af regning** {genudskrivning$_{v.nom}$ af$_{prep}$ regning$_{n.sg}$}	'reprinting of bill'	TDC (1)
7.25	**fakturaprint** {fakturaprint$_{n.sg}$}	'invoice print(ing)'	TDC (1)

The relationship between morphological form and "type" of query, or rather query intent, deserves to be investigated in much greater detail, as it is far from having been fully understood. We shall give a few suggestions as to the actual mapping between form and intent, but most of all we shall be contributing to a proper understanding of why users make the morphological and syntactic choices they do, and especially how they come to refer to a certain aspect of their need in the first place, due to a conviction that users' choices represent the link between form and intention.

7.2.3 Alternatives to the standard trichotomy

It is often forgotten that there were many alternative suggestions in papers predating Broder, which just never caught on. Jansen *et al.* (2008) review a range of the plethora of alternative classifications available and come up with the following generalized schema which, significantly, treats the Broderian trichotomy as fundamental[158].

Jansen *et al.* (2008)

1. **Informational**: Queries meant to obtain data or information in order to address an information need, desire, or curiosity
    a. **Directed**: Specific question
        i. **Closed**: Deals with one topic; question with one, unambiguous answer
        ii. **Open**: Deals with two or more topics
    b. **Undirected**: Tell me everything about a topic
    c. **List**: List of candidates
    d. **Find**: Locate where some real world service or product can be obtained
    e. **Advice**: Advice, ideas, suggestions, instructions

[158] See also Lee *et al.* (2005) and Daoud *et al.* (2009), among others.

2. **Navigational**: Queries looking for a specific URL
    a. **Navigation to transactional**: The URL the user wants is a transactional site
    b. **Navigation to informational**: The URL the user wants is an informational site
3. **Transactional**: Queries looking for resources that require another step to be useful
    a. **Obtain**: Obtain a specific resource or object
        i. **Online**: The resource will be obtained online
        ii. **Off-line**: The resource will be obtained off-line and may require additional actions by the user
    b. **Download**: Find a file to download
        i. **Free**: The downloadable file is free
        ii. **Not free**: The downloadable file is not necessarily free
    c. **Results page**: Obtain a resource that one can print, save, or read from the search engine results page
        i. **Links**: The resource appears in the title, summary or URL of one or more of the results on the search engine results page (As an example, a user enters the title of a conference paper in order to locate the page numbers, which usually appear in one or more of the results)
        ii. **Other**: The resource does not appear in one of the results but somewhere else on the search engine results page (As an example, a user enters a query term to check for spelling with no interest in the results listing)
    d. **Interact**: Interact with program/resource on another Website

Presumably, Broder's trichotomy survived not least because its extreme simplicity had a certain appeal for people who otherwise found it difficult to perceive much regularity or patterns in search behaviour. Whatever other reasons there might have been, researchers were probably right in turning to this new and simpler typology, because other suggestions typically focused more on why and how people used the search engine, thereby contributing more to social studies than to the development of superior search algorithms. For instance, Sellen *et al.* (2002) published a similar typology in the same year as Broder, but theirs had the following more behaviour-specific categories.

Sellen *et al.* (2002)

- **Finding**: Using the Web to find something specific. In this case, searching is goal-oriented and very well defined: E.g., Finding a fact such as a phone number, spelling or product name; a set of facts such as a list of ingredients for a recipe, or list of train times; or a virtual product or products such as a document, software, map, or image.
- **Information gathering**: Less specific than "Finding", but using the Web to purposefully research a specific topic for various reasons. E.g., Gathering information in order to compare, choose or decide about something (such as buying products or looking for jobs); in order to supplement a future task (such as collecting background information to write a document, or to prepare for a meeting); or in order to be inspired or get ideas.
- **Browsing**: Going to sites out of personal or work-related interest with no specific goal in mind but rather to be informed, stay up to date or be entertained. E.g., Browsing through a newspaper or magazine, following an interesting link, or checking to see what's new on a hobby-related site.
- **Transacting**: Using the Web to execute a transaction securing future products or services: E.g., Making a bank transfer, paying a bill, ordering a physical product, or filling out questionnaires.
- **Communicating**: Using the Web in order to participate in chatrooms or discussion groups.
- **Housekeeping**: Using the Web to check or maintain the accuracy and functionality of Web resources. E.g., Checking that information on a Web site is up to date, that links are working properly and so on.

In the present author's opinion, these alternative classifications tend to mix different criteria in unwarranted ways. As interesting as the above types are in themselves, they tell us more about why and how people use the internet than about the relationship between expression and need. Focusing on search behaviour and assigning various behaviours to labelled groups is certainly a reasonable first step towards producing a classification, but very little real understanding is achieved unless the analysis proceeds to the next step of identifying the different parameters along which the data vary, in addition to the observed values. There are certainly different behaviours corresponding to the various types, but to what extent and in what ways are the types defined and distinguished, and do they really belong in the same hierarchy? We have argued that they do not. Accordingly, this is not the kind of intention that we shall want to typologize over. Broder's types are more to the point in this sense, but there is still a long way to go before a really useful typology based on query form and the intention reflected in it is achieved.

Jansen *et al.* classify the various attempts at typologizing goals or intentions using one synthesizing typology. The idea is good, and their contribution is important, but in the end it hardly seems appropriate to include the various types in the same typology at all, as they actually classify entirely different phenomena. For instance, Sellen *et al.*'s *browsing* type cannot be properly equated with the navigational type as implied by the classification in Jansen *et al.*, because browsing is a form of behaviour on the external level, which may potentially involve navigational queries but in fact need not implicate search at all. Other types involve features of what we shall call *search types*, which are treated in section 7.3.2. An alternative that will be developed in later chapters is not to attempt to guess at user intentions at all, but rather to classify queries according to what the constitutive terms actually refer to. This minimizes the risk of biased classification of queries by being true to the queries as the ideal IR system would see them. According to the model developed below, reference can be made to the Aim or the Goal, or to one of six other basic referential targets involved in search.

7.3 Search and ye shall find

The term *information retrieval* was originally coined by the mathematician, physicist and IR pioneer Calvin Northtrup Mooers and was intended to cover "the intellectual aspects of the description of information and its specification for search, and also whatever systems, technique, or machines that are employed to carry out the operation" (Mooers, 1951, p. 25). It is noticeable that the word *retrieval* ordinarily denotes the recovery of something hidden or lost (see box *Retrieval*). In the case of IR, this must properly refer to data hidden in a depository consisting of a database or a collection of resources. The information carried by documents can hardly be said to be retrieved; rather, *documents* are retrieved so that users may extract (or rather, construct) the needed information from the data that they contain. IR is therefore a form of document retrieval (Ingwersen, 1992, p. 51)[159]. The term *information retrieval* has the unlucky connotation that the information is somehow stored in a database, just waiting to be uncovered. This in turn implies that the informational contents of a document are fixed and equal for all readers, which is known not to be the case; people interpret documents differently and construct different information from the same resources, depending on their background, mood and needs (Morris, 1994; Ingwersen & Järvelin, 2005a)[160].

Saracevic (2007) emphasizes that retrieval of data from databases (i.e., retrieval pure and simple) centres on *aboutness*, whereas search is about *relevance*. His point is that retrieval of data, such as traditional library systems, does not involve search (or relevance). Rather, such systems perform matching of query terms to topic labels only. There is no question of relevance because the topic label can never be relevant to the user; it can only signal aboutness. An alternative to this position would be to consider both to involve retrieval, adding that data retrieval also includes a matching step, while information retrieval requires a searching step.

It should be pointed out in this connection that the study of IR may also be suffering from the metaphor of *searching*, which – unlike *information retrieval* – is not a technical term but a widespread layman's notion that shapes the very concept of what IR is and should be. Admittedly, it is one of the best non-technical words available in the English language, but it inevitably generates some unfortunate connotations, which may mislead the researcher (and the user) conceptually. As discussed

Retrieval

Synonyms provided by Merriam-Webster's online dictionary: *Recapture, reclamation, recoupment, repossession, recovery*.

Specifically, in connection with information retrieval, the verb is explained as *to get and bring back, to recover from storage*.

Historically, *retrieve* comes from a predecessor of the French *retrouver*, i.e. to re-find.

Search

In Merriam-Webster's online Learner's Dictionary (2012), cases where the item searched is not known in advance are notably marked as figurative uses, cf. *scientists are searching for a cure*, which they paraphrase as *SCIENTISTS ARE DOING RESEARCH TO TRY AND FIND A CURE*. Significantly, the IR usage is assigned to a separate sense of the word. This implies that the authors neither consider the act of searching for information identical to (or even closely related to) the search for something lost, nor to be a figurative use of the word. In other words, the original meaning of the word *search* does not cover the case of IR.

[159] The term *data retrieval* has been used by some in a more restricted sense of an inferential system, i.e. one that is able to answer a question by following lines of inference (Cooper, 1971). That term should accordingly be avoided in this connection.

[160] There is only one sense in which information can be said to be retrieved and that is when the term is interpreted as meaning *information-as-thing*, i.e. as a synonym of *document*, see section 4.4.1.

at length by Sartre (1958), *searching* for something basically implies looking for an entity in an emptiness of not-that-entity and being able to recognize the item at first sight (see box *Search*)[161]. In fact, the reality of information retrieval is very different from this picture; in what is probably the majority of cases (albeit not all) search is not about finding something recognizable at all, but rather about sorting through a vast sea of information until an item is met with that promises to satisfy the user's vague and fluctuating needs. This is not normally a sense assigned to the verb *search* outside the context of IR. It *is* a sense associated with the word *hunt* (see box *Hunt*), which would have been a delightfully picturesque metaphor, but alas practice has been established a long time ago, and so *search* will have to do the job at present. We shall, however, be using the term *hunt* in characterizing one of the common subtypes of search to be defined in the following section.

> ***Hunt***
>
> Merriam-Webster's online Learner's Dictionary (2012) identifies the following senses of the word, among others:
>
> (Tr.) to traverse in search of prey; (Intr.) to attempt to find something. One of their examples seems to fit IR quite well: *She hunted around in the closet for a pair of shoes.*

7.3.1 Searching and finding

Based on Sartre's work, philosopher Flores Morador (2012) distinguishes between **searching** and **finding**, where the former implies looking for something known in an unfamiliar environment and the latter finding or stumbling on something unknown (but possibly suspected or previously described) in a familiar environment. Either the object sought or the environment must be known or familiar in some way in order that the viewer may realize that there is something there to find at all (Sartre, 1958). It seems to me that the familiarity in the case of *searching* where the sought entity is known is not material here and that the two possible types may thus be defined in the following way.

> <u>Searching</u>: Determining the possible presence of a known figure (on a ground which may be familiar or unfamiliar).

[161] The present author does not agree with Sartre that the nothingness representing the awareness of a lack of the sought entity is created by the viewer; what is subjective is the *relevance* of that nothingness to the searching person, cf. Part II. Thus, the user cannot be held responsible for the absence of any relevant material to retrieve on a certain topic but only for the sense of lack that this causes which is entirely conditioned by the user's perception of relevance and only indirectly related to real-world circumstances. This, however, is connected with the concept of information that one chooses to adopt; some authors feel that information recipients (constructors) are also information senders (Ingwersen & Järvelin, 2005a), as discussed in section 4.4.1.

<u>Finding</u>: Determining the possible presence of an unknown figure of relevance to the finder (on a familiar ground).

The familiarity of the ground is what makes the appearance of a totally unknown figure relevant and hence this aspect is covered by the main part of the definition. Relevance may, however, also be caused by the finder having received a description of the figure (Sartre, 1958; Flores Morador, 2012); in this case, matching of description and entity will take place. The question is how Flores Morador's two categories apply to information retrieval. In a real-world environment, it is necessary to scan a scene in *search* of a known entity, but in a digital environment this is a simple case of matching. It would therefore seem that *finding* is what most types of digital search are about (just as searching for a cure really means trying to *find* a cure; cf. box *Search*). However, things are far more complex than this because the sought object which may be known or unknown is not necessarily a substantial document; rather, it can be any *facet* of a resource (cf. section 2.3). It is possible to search for a known document, string, topic or piece of information without knowing where to find it, both as regards the resource containing it and the background of all available resources. In other words, looking for any document containing a certain string means *searching* (in the technical sense of Flores Morador) for a known formal resource and *finding* a substantial resource that features it, whereas looking for any document on a certain topic means searching for a known representational resource and finding an unknown formal and substantial resource which provide the information in question (fig. 7.1). Looking for anything which might be useful in a given situation means searching for an interpersonal facet and finding a relevant representational, formal and substantial resource. Even purely serendipitous search for entertaining resources may be interpreted as searching for a known effective facet and finding relevant material based on that[162]. Thus, in IR there is always simultaneous searching and finding going on (unless all facets are known in advance, as in prototypical refinding), and the crucial difference between the types is the facet that acts as the **target**

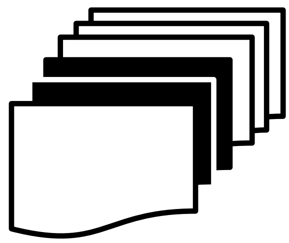


Fig. 7.1. Frequently, only one or two facets are known in advance when a search is performed, and it can be argued that these are sought while the others are merely found if something turns up which satisfies the need.

[162] The idea that one facet is distributed across many similar resources may seem highly abstract at first. However, consider a case in which *the same* resource is available as two identical files retrieved from different locations. In such instances we would not want to say that a user was searching for one rather than the other, and in the same way if you search for a certain effective resource, such as some kind of entertainment, then the known target is the effective resource that covers all documents with the appropriate characteristics; if the effects are identical from some user's point of view then these are basically irrelevant variations of *the same resource* as regards the effective facet.

of the search by being the known entity searched for (or *one of the known entities* if more than one facet is known in advance), in the sense that it is the one that the expression refers to[163]. The Faceted Stimulus model explains a lot of what goes on in search and can even be used for classifying search types, as we shall see in the following section. The terms *IR* and *search* are so deeply ingrained in the research field that it would be foolish to attempt to change them at this point, and we shall consequently stick to them. We shall also not uphold the distinction between *searching* and *finding* in the sense of Flores Morador (2012) as a dichotomy between types of performing information retrieval in the following since we have seen that both normally occur simultaneously in IR and that it is rather a question of which facets are sought and which found[164]. It is clear that the immense importance of facets deserves appreciation and that searches for different facets should not be treated alike. In the following, we shall define a range of basic search types which recognize the fundamental difference between searching for something known and entering a query in the hope of getting something – anything – relevant, which is crucial for the way users express themselves.

7.3.2 Search types

The overarching classification of goal types provided by Jansen *et al.* (2008) shown in section 7.2.3 above appears to mix *user intent classification* with what for want of a better name one might call **search types**. This confusion is inherited from the literature on which it is based and is natural given that it can be very difficult to discriminate among classification criteria. We shall attempt to disentangle the implicit taxonomies by first focusing on pure *search types* and then suggesting a novel way of classifying *user intents* in chapter 8. Blair (2002a; 2006) suggests that there are two types of search: *Exhaustive* and *sample searches*[165]. With an exhaustive search, the user wishes to identify *all* relevant resources, whereas with a sample search only a certain amount of material is required. In the same vein, but with a different outcome, we might tentatively suggest the following more encompassing

[163] It is possible to refer to more than one target in the same query, but the empirical data presented below show that it is more common to restrict oneself to one at a time.

[164] Indeed, this would appear to be true in general, even outside IR; all entities have facets when perceived by an observer, and when you have a description of the unknown entity you in fact search for the formal facet of it (the type) or sometimes the representational or interpersonal facet (something that is useful for something or in a given situation) and finding the substantial facet.

[165] Blair combines these types with two other criteria to produce a set of eight types, but these are of a different kind than the ones that we are calling *search types* here. One is the size of the resource collection, which may certainly influence user behaviour and system effectiveness, but is inherently of a different nature, and the other is the difference between *content* and *context searches*, i.e. between searching for the contents of documents as opposed to specifying document characteristics and metadata in order to *navigate* or *teleport* to the desired information. This is a difference that deals specifically with user behaviour in the form of the choices made when formulating a query, which is what we shall be looking at in the following chapters; it does not belong with search types, however.

typology of eight basic search types (note that the terminology used here does not match the one suggested by Blair).

1. **Hunting search**: Find anything
    a. **Sample search**: Samples / Occurrences
    b. **Definitive search**: Exhaustive list / Definition
2. **Retrieval search**: Find exact target
    a. **Specific search**: Find something specific
    b. **Collection search**: Find all instances of a type
3. **Refinding search**: Find something previously seen
    a. **Refind resource search**
    b. **Refind results search**
4. **Results search**: Check distribution / Existence
    a. **Frequency search**: Check existence / Correctness
    b. **Distribution search**: Check distribution

This simple classification avoids making any reference to the desired type of resource or what the user intends to do with it; it deliberately neutralizes the difference between data, information and knowledge and between documents and content; it includes no motivation and no goals, only an abstract classification of basic search types for which one is hard put to come up with a name. The reason for this is not that such other differences are unimportant – quite on the contrary – but they belong to different subsystems. Search types in the sense introduced here are of a mathematico-logical kind. The various search types are exemplified and described in depth below – the explanatory notes in the list above are only for fast reference.

⬤ 1a – Sample search

In this type of search, the user is merely after mentions or images (etc.) of instances of a certain type. In this case, the query *rodent* should return examples of various RODENTS. There would be no need for exhaustiveness. Thus, the following query is not likely to require an exhaustive listing.

7.26	**beer apps** {(beer $apps_{n.pl}$)}	TDC (12)

Rather, the user wishes to locate an entertaining app of the specified type. It is not necessarily the case that any app will do. Rather, the user does not know in advance which one (or indeed, if any) will suit his or her needs and desires. The same is true of sought information. A traditional informational search is of this type, provided that the user only requires the information to be useful in a given situation and not necessarily exhaustive. A typical example from L^{TDC} is the following.

7.27	**virus** {$virus_{n.sg}$}	TDC (1817)

This is most likely a request for information on viruses and what to do if your device is infected. It is less likely (though by no means inconceivable) that it is a request for a definition of the concept, and as it is in the singular the user is not likely to want an exhaustive list of known viruses either.

Most prototypically, sample search targets the representational facet (i.e., a topical search). However, it is possible to want just any formal resource with a certain string in it or any substantial resource with certain characteristics, etc. (see further section 7.3.3).

⬤ 1b – Definitive search

In this type of search, the user is interested in obtaining a definition of a concept, or an exhaustive list of all instances of that concept. A search for *rodent* should in this case result in definitional information on the class of RODENTS, and/or a list of all members. The query would be most likely to be rather simple, referring to a specific concept.

7.28	**ip telefoner** {ip $telefoner_{n.pl}$}	'ip telephones'	TDC (60)

This could be a request for information on what IP telephones are, but it may of course also have been a sample search for available products of this type.

Listing all members of a class and giving a definition of it may initially seem to be very different tasks, but this search type is so abstract that it covers both – it involves searching for something that describes the sought fully and completely. There is indeed an important difference between exhaustive lists and definitions, but it is not one of search type; it is one of reference. In ex. 7.28, the term *ip telefoner* refers to the concept of IP TELEPHONES rather than the individual objects, and consequently it is the concept that requires an exhaustive description rather than the class of entities.

Combinations of concepts are also possible, such as the following[166].

7.29	**nokia telefoner** {(nokia$_{n.prop}$) (telefoner$_{n.pl}$)}	'nokia telephones'	TDC (47)

This is a cross-section of the concepts NOKIA and TELEPHONES. If the user intended this query to result in an overview of available models, then this would be an example of the exhaustive list type. It might of course also be a request for information on Nokia phones in general, which could be a case either of sample or definitive search.

Included in the exhaustive category is also the negative type encountered in situations such as patent search, where it is absolutely crucial that no instance exists in the collection; the search has to be exhaustive in order for this to be possible. Blair (2002a; 2006) treats such *existence searches* as a subtype of exhaustive search.

● 2a – Specific search

In this type of search, the user is looking for something particular that is already known or at least assumed to exist, such as a specific mention of rodents in the media, or a video that someone has told the user about. The user would be likely to attempt to predict the phrasing of the particular occurrence, or the title or description of the video, rather than describing his or her situation of need, for instance, but this is just typical behaviour and not part of the definition of the type.

The *specific search* type also covers *fact finding*, because of the absence of a distinction at this level between documents and information (and other facets); a fact is a specific piece of information which can be pinpointed in the same way as a specific string, even though the

[166] There is no clear boundary between cases with one concept to which reference is made with a nominal compound and cases where two concepts are combined; both ex. 7.28 and 7.29 may be interpreted either way.

contents of the factual information need not be known, just as the meaning of an anticipated string is not necessarily understood.

● 2b – Collection search

This type of search is somewhat similar to type 1b in that the user wants to retrieve all that there is of something, whether it be information on a given topic or instances of a certain type. For instance, the user might for some reason be interested in finding all articles on rodents in a certain newspaper during the past year. The difference is that here the point is not to find something that covers the topic completely, but rather to retrieve all available resources on the topic in a particular environment. This type is used when researching the literature on a topic rather than the topic itself, a task that is common in scientific research, for instance.

● 3a – Refind resource

This type of search is related to type 2a in that a resource that is known to exist is sought again, whether a page, document or other file. The user would be likely to either reuse the query that was issued the previous time the resource was found or match a particular stretch of text on the site, rather than describe the situation of need or even the kind of resource needed. In spite of their similarity, this is not a subtype of 2a, because what is refound is the resource rather than the information it contains. In fact, type 3a also covers cases where users do not actually know that the resource exists but merely think that it might and take their chance, which is one interpretation of navigational search.

● 3b – Refind results

In this type of search, the user is interested in getting the exact same results as on a previous occasion. In such cases, the user is most likely to reissue the same query again. Receiving the exact same results a second time is often not possible in a modern personalized IR system – especially not if some time has passed since the last search – because the results will change every time a query is issued as a result of the user profile being updated.

● 4a – Frequency search

It is very common today for people to use a web search engine such as Google for checking the accepted spelling of a certain term or the correctness of an expression. Often this involves comparing the frequency of several variants in use and the reliability of the sources. This type is not the same as the one called *existence search* by Blair (2002a), as mentioned above. The sense intended here is the kind of search where the user is merely interested in knowing whether something occurs but does not require there to be no instance, i.e. exhaustivity does not enter into it.

● 4b – Distribution search

Sometimes studying the distribution of a term or construction is the primary goal. This might be the case if a user wanted to compare the number of mentions of women to that of men on a certain site, for instance. In other words, these users are not interested in the information as such, but merely its distribution.

The definitions above contain several hints that there is again a close relationship with facets, and indeed most of the search types can in fact alternatively be defined on the basis of facets, which furthermore provide a much more fine-grained typology. Consequently, at the risk of being guilty of what Swanson (1988) dubbed "intellectual skeet shooting"[167], we shall proceed to take the typology of search types apart and reconstruct it on the basis of differences in targeted facets. The six facets of a *resource* cannot replace those distinctions, but the facets of topic and information which were briefly suggested in section 4.5.4 can. The typology of search types is intuitive and captures behaviour in typical cases and is useful for these reasons, but in order to arrive at a fuller understanding of that behaviour we shall need – for our purposes – to look at it from the point of view of facets.

7.3.3 Topic and information facets

Apart from targeting one of the six facets of resources and the information that they hold, queries may also in themselves represent various facets of a topic in their role as topic labels, and it is possible to target these facets rather than the actual resources sought. In fact, a prototypical topical search does not usually target the representational facet of a resource, unless the user is interested in finding all resources on a certain topic rather than sufficient information on the topic; instead, it typically targets the representational facet of the information as represented by the topic label. We shall call queries that target a topic facet **Subject-targeting** and those that target an information facet **Object-targeting**. The difference is in whether it is the resources treating the topic or the topic domain that is the target and which the user wants information about. The following types occur.

[167] Cf. "Bar Hillel seemed to enjoy intellectual skeet shooting; he would often toss out good ideas and then blast them to bits. Others would pick up the pieces and find them interesting" (Swanson, 1988, p. 93).

1. Subject-targeting types
    a. **Substantial topic**: Tell me about this string (rare)
    b. **Formal topic**: Tell me about this word or expression
    c. **Representational topic**: Tell me about this topic (domain)
    d. **Interpersonal topic**: Give me this particular information (fact)
2. Object-targeting types
    a. **Intentional information**: Show me resources intended for me (rarely expressed as a query)
    b. **Substantial information**: Show me resources with this string in them
    c. **Formal information**: Show me resources with this expression in them (in this language)
    d. **Representational information**: Show me resources with information on this topic
    e. **Interpersonal information**: Show me resources that provide the information that I need
    f. **Effective information**: Show me resources that affect me emotionally in the desired way

Perhaps surprisingly, facets even account for the difference between exhaustive and sample searches, because you can only expect a search to be exhaustive if it targets the information rather than a topic; you cannot reasonably ask for all *possible* information on a topic but only for all *available* information, and consequently Subject-targeting search is never exhaustive. On the other hand, it is of course possible to only want sample information in an Object-targeting search, but this is not in any way part of the query; it is merely a behavioural observation. It is an interesting fact on which the system would do well to collect information, but it would have to be collected separately. The various types are explained more thoroughly below.

● **Substantial topic searches**

A substantial topic search is employed for checking whether a string exists or occurs frequently. The user is searching for a string in order to acquire information about the string rather than with the purpose of retrieving documents. Thus, the results list will typically be enough in the way of a response, and no resource need be clicked. The string need not represent a linguistic expression and consequently no linguistic alternatives could be suggested for improved search.

● ***Formal topic searches***

A formal topic search is employed for checking whether a word exists or is correctly spelled, and whether it is in frequent use. Also, the user may want to find out what the term means (i.e., how it is defined). The user is searching for a term in order to acquire information about the term rather than with the purpose of retrieving documents. Thus, the results list will typically be enough in the way of a response, and no resource need be clicked. The string must represent a linguistic expression. Suggestions for improved search might include synonyms, but not related terms.

● **Representational topic searches**

A representational topic search refers to a topic that the user wants to understand better. This time it is not merely a matter of finding a definition of the term but of understanding the broader topic connected with it (or with a combination of terms). The string should represent a topic, i.e. it would not normally be a verb or an adjective, though nominalizations are certainly possible. Suggestions for improved search might include related terms that belong to the same domain.

● **Interpersonal topic searches**

An interpersonal topic search is performed when it is known with relative certainty what information is needed, such as is the case with fact-finding.

● **Intentional information searches**

The intentional information search type represents the case where users want the information intended for them. This is not normally something that is expressed as a query, though it might potentially be.

● **Substantial information searches**

With a substantial information search, the purpose is to find the documents in which a string occurs. This target is likely to be used if a resource has been seen before and is sought again by issuing a query containing a string from that resource. The string need not represent a linguistic expression[168].

- **Formal information searches**

With a formal information search, the point is to locate documents that mention a certain term. The user's need is thus not to understand the term but rather to find the documents mentioning it. The string must represent a linguistic expression. One purpose could be finding recent mentions in the media of some topic, or relevant literature on the topic. In this case, however, targeting the representational facet is more to the point.

- **Representational information searches**

With a representational information search, the purpose is to locate documents that treat a certain subject (rather than merely mentioning a term, as in a formal information search). One purpose could be finding recent mentions in the media of some topic, or relevant literature on the topic. The difference from formal information searches is only slight, but with representational information searches the string must represent a topic, i.e. it would not normally be a verb or an adjective, though nominalizations are certainly possible.

- **Interpersonal information searches**

An interpersonal information search is one that looks for resources which provide the exact information that the user needs. This of course requires the user to know what is needed. Interpersonal information searches would also be used if people were looking for alternative solutions, such as different sellers of the same type of product or providers of the same kind of service.

- **Effective information searches**

The prototypical effective information search is the serendipity search for entertainment.

According to the framework suggested here, many searches which are not usually considered topical do in fact belong in that category. This is because the notion of topic has been extended here to cover all the different facets, from the actual string (the topic label) over the linguistic terms and topic domains to situations involving an instance of the type in question. Even actions and events can be topics; they then typically occur in infinitive, participial or nominalized form. It seems natural to suppose that nominalizations tend to represent representational topics (being more similar to prototypical topic labels due to their nominal characteristics), whereas infinitives and participles describe real or potential

[168] In principle, it may even be an image supplied by the user.

situations involving such an activity with a view to getting advice or information regarding it. The exact distribution is not known, however.

Subject- and Object-targeting queries reflect the basic distinctions of the Stage-based Intent model developed in chapter 8, where even more types are added in accordance with the enormous amount of variation that can be observed in query logs. Unfortunately, even though facets are an integral part of queries and therefore at the centre of our interest here, what facet a query targets is by definition not recoverable from the form of the query. After all, any expression has all of these facets, and it is not possible to determine from a string alone what facet was targeted. This, then, is something else on which a LAT designer might consider collecting information during a search. Even so, educated guesses may sometimes be possible on the basis of query semantics, and even with reference to the choice of morphological form and syntax. Alternatively, several sets of responses may be presented to the user, corresponding to various interpretations of the inherently multifaceted query.

8 Making sense of a query

> *I focus instead on what the theory can contribute to a general framework [...]. When I work in this manner, scholars who prefer more specific and applied approaches are often dissatisfied, though at the same time the philosophical inconsistencies in the theory's framework do not trouble them much.*
>
> (Brier, 2008)

All classification of queries is based on assumed or professed user intentions. However, classifying queries according to user intentions or needs in a satisfactory manner is a daunting task; how would one even begin to describe the characteristics of different intentions or decide what features to focus on in the first place? How would one describe the ways in which intentions, needs or problematic situations may possibly differ? Even if one were to succeed in this, how would one decide where to draw the line between intentions different enough to merit separate treatment? Deciding which intentions are "the same" or even just similar and which ones are different is a serious challenge.

The query characteristics that state-of-the-art query classification pays attention to are ones that echo the user's intention indirectly; it seems to be implicitly assumed that there is no *direct* reflection of the user's intention in the expression apart from the lexical meaning of the words used. However, the way we express our needs and phrase our queries may indeed signal our intentions more directly, just as with other uses of language. It is inevitable that people should make lexical, morphological and syntactic choices that correspond to the needs they want to communicate and that they should expect their interlocutors to be able to decode their intentions from the clues that such choices provide; yet computers are at present largely unable to perform this interpretation. When searching, people are presumably well aware that the addressee is not human and has limited capabilities, but even so they still do not communicate their needs in the form of isolated references to concepts. Rather, they use some form of language. At times this language is rather fragmented and messy, to be sure, but a form of language nonetheless.

If we are to make any real progress, we must be open to the idea that more is communicated in a query than the subject on which the user desires information. On the other hand, we must also acknowledge that we can only ever capture intentions that are somehow mirrored in queries, however faint the traces may be in some cases. In other words, we shall need a definition of **intent** that differs from the normal understanding of **intention**; one that makes reference to the query.

> Intent: A user intention as expressed in a query through the expressive choices made by the user, as understood in the context in which the query was produced.

What signals the user's intent is in fact the *choices* made by the user when deciding how to phrase the query; it is when there was a choice that the form of the query can tell us something. If the user could have used either an infinitive or an imperative, then the choice signals user intent and it becomes interesting to find out what exactly the difference might indicate[169], and if otherwise identical queries occur sometimes with a singular and sometimes with a plural term then it is natural to wonder what difference in intention this choice reflects.

As opposed to intentions which are not easily disentangled, *intents* can be analysed into a range of types. This is because they depend strictly on what can be conveyed through a query and as such they are a far more manageable set than full-fledged intentions with all their inherent abstractness, vagueness and ambiguity. However, there is still a whole range of possible ways in which intents might possibly be classified. Consequently, different attempts at classification are still not entirely comparable.

The tendency for queries to reflect intentions is both reinforced and masked by the fact that what people are looking for is typically stretches of human language as well. Consequently, search queries frequently incorporate references to the (often hypothetical or **anticipated**) text the user is hoping to find. If a query consists of a verb in the past or present tense it is most likely a case of anticipated text[170]. Specifically, a query such as the following, which consists of the present form of a verb, is likely to be an example of what we shall call an **anticipated header**.

8.1	**tilslutter** {tilslutter$_{v.pres}$}	'connect(s)'	TDC (29)

[169] This particular case is investigated in section 9.4.
[170] Appendix D1 contains a list of such queries in L^{TDC}.

Most likely the users who issued this query anticipated a header or an instruction with this word in it – something akin to *SÅDAN TILSLUTTER DU*..., i.e. 'This is how you connect...'. Had they merely been looking for information on the topic of connecting something, they would be expected to use a less marked verb form than the present, such as the infinitive *tilslutte* or the nominalization *tilslutning*, as indeed far more people do.

8.2	**tilslutning** {tilslutning$_{v.nom}$}	'connection'	TDC (490)

It may of course very well be the case that these users have the same ultimate Goal, but they decide to go about getting there in different ways; they make different choices, and this is likely to be linked to their mental treatment of the problem at hand and the process of growing understanding and planning that they are undergoing while searching. Thus, the chosen form may potentially tell us a lot about the user's plans and hopes, and about his or her certainty or uncertainty as regards the immediate Aim and current ideas about the best way of solving the problem that triggered the search. It does not necessarily tell us much about the kind of resource that the user is imagining would satisfy the need, but then the user may not know what resource would fill this role. It is all too easily taken for granted that most or even all users search for anticipated text in resources, this being the original primitive way of conducting a search. It is true that this kind of search is still common, but it is far from being the prevalent type in all domains. We shall see, for instance, that in L^{WID} most users search for their Goals, typically in the form of a job description, a work task or something they would like to work with, rather than mere strings to be matched. A few examples will suffice for illustration[171].

8.3	**cleaner** {Goal:JobTitle: cleaner$_{n.sg}$}	WID (491)
8.4	**cleaning** {Goal:WorkTask: cleaning$_{v.nom}$}	WID (1156)
8.5	**clean** {Goal:WorkTask: clean$_{v.inf}$}	WID (159)

[171] The labels used in this particular analysis are explained in section 10.1.

8.6	**driver** {Goal:JobTitle: driver$_{n.sg}$}	WID (1205)
8.7	**driving** {Goal:WorkTask: driving$_{v.nom}$}	WID (86)
8.8	**drive** {Goal:WorkTask: drive$_{v.inf}$}	WID (13)

8.9	**milker** {Goal:JobTitle: milker$_{n.sg}$}	WID (7)
8.10	**milking** {Goal:WorkTask: milking$_{v.nom}$}	WID (130)
8.11	**milk** {Goal:WorkTask: milk$_{v.inf}$} / {Goal:Core: milk$_{n.sg}$}	WID (17)
8.12	**milking cows** {Goal:WorkTask: milking$_{v.nom}$ cows$_{n.pl}$}	WID (8)
8.13	**milk farm** {WorkPlace:Environment: milk farm$_{n.sg}$}	WID (9)

Responding successfully to such queries goes beyond matching. Of course, they may in principle be interpreted as instances of anticipated text just like the tensed forms, assuming that users are looking for ideal documents containing certain strings. Indeed, it is always possible to argue of any query that it represents a mere string search in this way, and in fact all queries probably *are* string searches, but crucially most of them are not *only* that. Just as we are not bound by our senses to view a certain situation in one way at a time but can simultaneously conceptualize a flock of sheep as an individualized mass consisting of many parts and as a collective of individualized sheep, we can also (consciously or subconsciously) intend a query to be simultaneously a string search and a Goal search, for instance. Thus, by searching for a job title we may be hoping both to find information that will help us get that job and to locate an actual job posting in the form of a digital document containing the string (or a synonym, if the system is known to support semantic extensions), and this divided intention is likely to influence user behaviour, especially as searchers are known to be highly insecure about their needs and the rationalities behind their search, as shown by Kuhlthau (1993), among others. It is likely that the simultaneous intentions take place on separate cognitive levels and that rational thought is involved to a varying extent in the different cases, as searching for an actual string is a conscious act that follows from rational planning of the search while searching for the ultimate Goal is more directly guided by the sense of lack or desire created by the situation that instigated the search and nourished by the uncertainty and anxiety generated in the process. It may well be that, if asked, users who search for their actual Goals rather than anticipated text would deny that there would be any sense in hoping for anything but exact matches, yet unconsciously they would be doing just that.

With current search tools, the rational thing to do is indeed to search for strings (as opposed to actual Goals), but normally the user has no way of knowing in advance what documents are available and what text they contain, and so they are forced to guess. If tools could be made to listen to the other kinds of search, the result would probably be more effective, given that users are more likely to be able to make an accurate estimate of what they need than guessing the exact wordings of unknown documents, in spite of the difficulties connected with recognizing an unarticulated need.

Interpreting a query as a string search when it might alternatively be classified in a different way impedes further progress when it comes to constructing helpful tools, because if the ultimate result is considered to be a document that matches the query string exactly, no further help can be given and that is the end of it. Thus, it is vital that all cases that can possibly be understood as more than just a mere case of anticipated text be specified as such and treated according to their most likely other interpretation. Consequently, only those queries which are particularly unlikely to be anything but anticipatory due to their form are classed as such in this thesis. The remaining queries are classified according to a framework developed below which comprises eight different targets, apart from the facets discussed above.

8.1 Referring to a need

Prior to phrasing a query, users pass through a stage of having to decide – often subconsciously – what to *refer* to in the first place. This is intimately connected with the types of task that they are facing, and is crucial to the way search is exploited and queries formulated. Due to the special setup of search communication, *reference* in the wide sense in which it is used here covers two general linguistic categories of phenomena. One is *reference sensu stricto* as traditionally understood, including *referentiality*, and the other comprises phenomena that linguists would normally characterize as instances of *mood* and *aspect*[172]. We shall concentrate on the latter types here.

Mood is a category that deals with the difference between referring to something that is currently true (but not necessarily occurring in the present), something that is desired or hoped for, something that the speaker wants someone (in this case typically the system interacted with) to perform, etc. In natural language, typical modal categories are **indicative** for what currently holds, **optative** for what is desired, **imperative** for the command, and **interrogative** for the question.

Aspect is many things, but for our purposes it is particularly relevant in two situations. One is to do with the difference between something that is true of the current state and something that will turn out to have been true from the point of view of some potential future situation, and the other is about the difference between what the future state of affairs will be like and what it looks like from the initial point of view where it represents a need. These aspectual categories are similar to some in natural language like the **retrospective** perfect for what happened in the past but still characterizes the subject in the present, and the **prospective** future for what is predicted about the future given the present conditions, as exemplified below (Dik, 1997, p. 238f).

8.14	**I visited New York** Past tense	8.15	**I have visited New York** Retrospective aspect
8.16	**One day, stocks will rise again** Future tense	8.17	**Stocks are going to rise again** Prospective aspect

Strictly speaking, all of these examples represent a tense *and* an aspect, but these two categories interact in ways that make the above descriptions entirely legitimate when dealing with English, where the two systems are intertwined. Ex. 8.14 and 8.15 are both literally about the past, but in fact the latter tells us more about the present; it is a statement about the current state of the speaker rather than about what happened at one time. Similarly, while ex. 8.16 is about the future, ex. 8.17 characterizes the present conditions by stating the way things are likely to develop based on those conditions.

[172] The reason for adopting this broader concept of reference was given in chapter 5.

8.1.1 The stages of search

The categories of mood and aspect are of great importance in search, because the whole point of searching is to get from one state to another – from a problematic situation (or at any rate one characterized by a lack, need or desire) to one of satisfaction, or at least a little more knowledge on how to proceed. From the point of view of real-world accomplishments, a search thus represents an attempt to transform an existing situation, which we shall be calling **Stage 1**, into a more desirable situation, **Stage 2** (fig. 8.1).

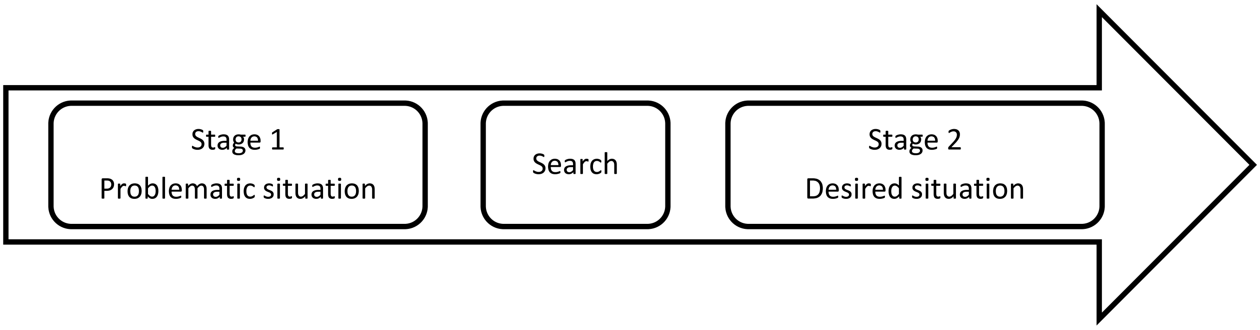


Fig. 8.1. Queries may refer either to Stage 1, the existing problematic situation, or to Stage 2, the ideal outcome.

The search may be initiated with a view to proceeding from the problematic situation, and this requires the user to decide (more or less consciously) on what to search for. Users may either try to describe their predicament in the hope of finding a solution (*Source-targeting*) or aim directly at the end state that they want to reach – one where the problem has been solved – and hope to receive advice on how to get there (*Goal-targeting*).

The change from Stage 1 to Stage 2 can be viewed from various points of view, and is accompanied by expectations and desires as well as being realized through the making of a request. The notions of mood and aspect describe the ways in which linguistically portrayed situations differ from real situations and the temporal angle and scale that define their portrayal. This is exactly what happens in queries; users portray situations which are either present problematic states or future wished-for situations. Actual modal and aspectual marking is not common in queries, except that both imperatives and interrogatives are surprisingly abundant, but the idea of modal and aspectual categories greatly helps the analyst understand the differences between various kinds of requests such as describing the state you are in from your own point of view as opposed to portraying it from the point of view of some desired ideal end state, and this is a point where searches differ markedly.

Targeting Stage 2 presupposes some initial idea of what direction you are going to move in, i.e. how the problem might possibly be solved, or at least what the desired end result would be like. Faced with the problem of having spilt red wine on wool, for example, it is possible to describe this problem, search for information on the topic of removing such spots, anticipate texts to this effect, target the actual removal in the hope of receiving information on how to carry it out, or describe the Goal state wished for.

8.18	**red wine on wool** {red wine$_{n.sg}$ on$_{prep}$ wool$_{n.sg}$}	*
8.19	**removing red wine stain** {removing$_{v.nom}$ red wine stain$_{n.sg}$}	*
8.20	**the following procedure will remove a red wine stain** {"the following procedure will remove a red wine stain"}	*
8.21	**remove red wine stain** {remove$_{v.imp}$ red wine stain$_{n.sg}$}	*
8.22	**spotless wool** {spotless wool$_{n.sg}$}	*

All of these solutions are met with in search logs. A person who has trouble turning on the sound on an iPhone may also deal with this problem in several ways, as in the following examples from L^{TDC}.

8.23	**iphone problemer** {(iphone$_{n.sg}$) (problemer$_{n.pl}$)}	'iphone problems'	TDC (3)
8.24	**iphone ingen lyd** {(iphone$_{n.sg}$) (ingen lyd$_{n.sg}$)}	'iphone no sound'	TDC (1)
8.25	**jeg mangler lyd** {jeg$_{n.pron}$ mangler$_{v.pres}$ lyd$_{n.sg}$}	'I lack sound'	TDC (1)

8.26	**iphone lyd** **Iphonelyd** {(iphone$_{n.sg}$) (lyd$_{n.sg}$)}	'iphone sound'	* TDC (1)

Stage 1 is not always as problematic as these examples may suggest; a search may be instigated by a much less trying situation if there is at the same time an idea of a (currently) more ideal Stage 2.

> The word 'problem' often seems too elaborate and dignified to denote what happens in minor cases of reflection. But in every case where reflective activity ensues, there is a process of intellectualizing what at first is merely an emotional quality of the whole situation. This conversion is effected by noting more definitely the conditions that constitute the trouble and cause the stoppage of action. (Dewey, 1933, p. 108)

For instance, many users in the Workindenmark job portal enter their nationality as a query; not because this represents their problem, but because it is the initial state, Stage 1, from which they wish to proceed to a Stage 2 which includes a new job.

Interestingly, a clear difference in search behaviour can be observed between searches from a problematic Stage 1 to a more normal unproblematic Stage 2 (HOUSE ON FIRE > HOUSE NOT ON FIRE; BROKEN USB > REPAIRED USB) and those that represent a progress from Stage 1 to a different Stage 2 which is not the normal situation in that sense, such as buying a product or installing additional features. After all, referring to Stage 2 requires – presumably for purely linguistic reasons – that the desired end state be not simply the normal default state that is so expected as to not have a name of its own, such as a house that is no longer on fire.

8.27	**house on fire** {house$_{n.sg}$ on$_{prep}$ fire$_{n.sg}$}	*
8.28	**extinguishing a fire** {extinguishing$_{v.nom}$ a fire$_{n.sg}$}	*
8.29	***fireless/*non-burning house** {fireless house$_{n.sg}$}	*

As a consequence, when there is a real problem users tend to describe Stage 1 (or an intermediate stage), whereas other desired changes require reference to the end state, or the effectuation of the change itself; they do not normally target Stage 1 because it is a normal state which does not imply any necessary action, let alone a solution. The exception is a LAT such as Workindenmark where it is part of the basic expectation that users want a job and hence the problem is already implicitly specified in context. This frees the user to provide other information than the nature of the problem to be solved. The Workindenmark case is studied in depth in section 10.1.

8.1.2 From problem to need

The process from the initial perception of a problematic situation or a felt need for information to the formulation of a query was discussed at some length in chapter 2. However, no real distinction was made there between problems (or problematic situations) and needs (or rather, situations characterized by needs). A problem is what causes the need, and as such the need should intuitively be directed toward something with some kind of mental substance such as a solution or a resource which might contribute to one, whereas a problem may not necessarily lead the user to perceive any solution. Thus, a need may – but need not – derive from a problematic situation.

Some information needs are closely related to, and completely dependent on, the problematic situation.

8.30	**tabt mobil i vand** {tabt$_{v.res}$ mobil$_{n.sg}$ i$_{prep}$ vand$_{n.sg}$}	‘dropped mobile (phone) in water’	TDC (3)
8.31	**reparation af mobil** {reparation$_{v.nom}$ af$_{prep}$ mobil$_{n.sg}$}	‘repair of mobile (phone)’	TDC (34)

The first of these examples describes the problematic situation directly and represents a plea for information on how to deal with it, while the second involves the recognition of a more particular need for information on a certain topic with a view to applying the information to the problematic situation. In such cases, the information need is clearly distinct from the problem that triggered the search (Ingwersen, 1992, p. 111). Sometimes, the need and the problem are so closely related as to be completely intertwined.

8.32	**find butik** {$find_{v.imp}$ $butik_{n.sg}$}	'locate shop'	TDC (905)

Here, the user's need is exactly that which is expressed, viz. identifying the whereabouts of a (TDC) shop. Other needs are more derived and abstract, but they can still be said to spring from a problematic situation in a wide sense.

8.33	**priser udland** {($priser_{n.pl}$) ($udland_{n.sg}$)}	'prices abroad'	TDC (797)

In this query, the link to the actual need-generating situation is somewhat less evident, though clearly still there in the background. Problems and needs must therefore be dealt with separately. The process of coming to an understanding of one's own problems and recognizing what needs this entails is a difficult one, and depending on users' ability to do so in a given case they may settle on describing the problem as best they can or try to search for something particular that they think would help them. To a user, the problematic situation involves a sense of lack, or "an inadequacy in what we may call his 'state of readiness' to interact purposefully with the world around him" (Mackay, 1960). In the tradition that goes back to Belkin (1980), the problematic state is seen as an *anomalous state of knowledge* (ASK) that needs to be brought back into harmony. This discord is the *mental trigger* that starts the process of establishing a need that can be satisfied through the interaction with a given tool, a process which must necessarily be part of, and central to, any model of query construction. There is no doubt that the degree to which users are able to identify their needs plays an important role in the choices they make when formulating queries, and also that their level of uncertainty affects what kind of information or help they would benefit most directly from receiving. We shall therefore review Taylor's (1968) four stages of need development as previously discussed in section 2.2 and see how they can be reinterpreted in terms of query expression.

- **The visceral need**

The *visceral need* occurs at a stage at which it is not yet known to the user what is needed. The sense of lack originates in emotional qualities such as a need for entertainment or greater confidence, etc. It may be possible at this point to describe the emotional urge, but not what the particular kind of resource that would satisfy the user would be like. The visceral need is both an early stage in the need development process and a kind of need to which the user may potentially refer in a query. The latter occurs but rarely, and admittedly is at the very limits of what one would expect to encounter, but it is at least conceivable that someone might search using queries such as *entertain me* or *information*. This is only likely to happen in the rare cases where a user never gets beyond the visceral need and hence is not able to formulate any more precise query. When it does occur, we shall call the entity referred to the **Motivation** behind the search. This provides us with the extreme limits that delineate the area within which all query reference must take place. We shall define Motivation in the following way.

Motivation: The emotional state characterized by a sense of lack that causes the user to approach a LAT in the first place.

Motivation is the first possible target in the Stage-based Intent model to develop in the user's mind. It is not, however, the first referential target that a user would think of referring to in a query, because the conscious process of expression deals with the targets in a different order from the development of the need, cf. section 8.3.

● **The conscious need**

The next stage in Taylor's model is the *conscious need* which develops in users' minds as they study the situation which they perceive to be problematic in some way or other and in which they find themselves. The conscious need is closely tied to the perception of the situation that caused the need, and corresponds to what we shall be calling the **Source** of the need (the situation that is the basis of the user's *Goal*)[173]. Again, as opposed to Taylor's notion, this new term will be used of the abstract entity that users may refer to through a query, the *target*.

Source: The situation that a user experiences and which guides search behaviour towards a solution to any problems associated with that situation or from which the user wishes to proceed to a more desirable situation.

Many users never get beyond the conscious need, or at least choose to reflect the Source situation in the queries that they produce. Perhaps this can be interpreted as an inability to come up with a more precise request for information, but on the other hand describing the problematic situation may simply be a conscious choice in some cases. Psychological experiments would be needed in order to get closer to an understanding of this behaviour.

[173] Note that this concept is totally distinct from the use of *source* in the discussion of information sources.

8.34	**har glemt min pinkode** ‘(I) have forgotten my pincode’ {$\text{har}_{\text{v.pres}}$ $\text{glemt}_{\text{v.perf}}$ $\text{min}_{\text{poss.1sg}}$ $\text{pinkode}_{\text{n.sg}}$}	TDC (3)

The users behind ex. 8.34 were probably not sure about what to do, so they stated their problem in the hope of finding a solution. Others in a similar situation opt to search directly for a new pincode (8.35) or refer to the action of producing a new code (8.36).

8.35	**ny pinkode** ‘new pincode’ {ny $\text{pinkode}_{\text{n.sg}}$}	TDC (84)
8.36	**lav ny pinkode** ‘make new pincode’ {$\text{lav}_{\text{v.imp}}$ ny $\text{pinkode}_{\text{n.sg}}$}	TDC (1)

These latter strategies target other stages in the process of understanding the problem and its solution – stages which we shall study in detail below.

● The formalized need

The third stage is the *formalized need*. At this point, users are able to point to the kind of response that they want from the system. Some users need answers to questions, others need information on a topic, some need access to certain resources and others want the system to perform an action or want help with performing one. When queries reflect this level of awareness (but not yet the exact identity of the object sought), the abstract entity of reference shall be called the **Trigger** of the search; it represents the need that triggered (the conscious part of) the search.

8.37	**hjælp til pc-problemer** ‘help with pc problems’ {$\text{hjælp}_{\text{v.nom}}$ til_{prep} $\text{pc-problemer}_{\text{n.pl}}$}	TDC (1)

This particular query does not refer to the contents of a resource in the same way that a topical query does not. Just as a topical query in fact refers to a predetermined topic in the user's mind, the Trigger-targeting query refers to the kind of help needed in the form in which it is represented in the user's mind rather than the material which will help gratify that need.

Trigger: The type of need experienced by the user, such as a need for a fact, information, resources, an answer, advice or (help with) the performance of an action.

The Trigger is often reflected in the morphosyntactic form of queries, such as interrogatives, topic labels or imperatives. Questions signal the need for an answer, while imperatives are requests for action, and topic labels are prototypically linked to the desire for information. The actual semantic content of the terms involved signal other targets, however, as will become clear below. This is a difficult aspect of Target Analysis, and recognizing Triggers takes some training.

● The compromised need

The fourth and last stage in Taylor's model is the *compromised need*. At this point in the process of need development, users are conscious of the difficulties involved in searching, and they incorporate this in their deliberations on how to perform an optimal search which will satisfy their needs. At this stage, people do not search for help with things that cannot be expected of a search engine. Instead, problems which need physical intervention are turned into information searches by using relevant topic labels. Users may then perform the necessary actions themselves after having acquired the needed information. Not everything that happens on this level is entirely deliberate, though; the so-called *label effect* applies at this stage, turning a large proportion of queries into topic labels due to a natural human propensity, for better or worse (Ingwersen, 1982; 1992, p. 116; Ingwersen & Järvelin, 2005a, p. 296ff).

When the compromised need is reflected in queries, we shall call the abstract entity referred to the **Subject**, reflecting their topical nature. Prototypical topical search is of this type.

Subject: The topic on which the user wants information, with its various facets.

> As shown in the previous chapter, Subjects have facets and one or more of these may be targeted by a term. This applies to the Source situation as well, though not all facets of a situation are equally likely or easy to express in words[174]. Most frequently, users portray actual situations with consequences for themselves (the interpersonal facet), or generalized situations of a certain type (the representational facet), and the latter is closely related to topical search (targeting the representational facet of the Subject) from which it can hardly be distinguished in a surface expression. There is an important difference in conceptualization, though: One is the portrayal of a type of situation that you may experience, and the other is a reference to a topic on which you may acquire information.

Experimental evidence confirms that Taylor's four stages of need development really do occur (Ingwersen, 1992). The empirical analyses of query logs performed below furthermore show the effects of the stages on query formulation. The Motivation, Source, Trigger and Subject are all Stage 1 targets; four additional targets belonging to Stage 2 are presented below, where they are discussed in detail.

8.1.3 User Goals and Aims

There can be various reasons for wanting to distinguish query types in the first place. One might be to identify the kind of content a user is looking for, another to find out *why* the user is looking for it, in order that he or she may be helped even more. These differences in reasons for constructing a typology of queries might seem innocent enough at first, but they actually result in highly different typologies, because the type of content sought does not correlate in any neat way with the ultimate Goal behind a search; a navigational search may locate a website through its name, but the Goal may be either to locate information or to buy products, or the search may be triggered by something else, such as a simple craving for entertainment. At the same time, information may indeed be located by searching for topics, but it may also be retrieved from dynamic sites that require interaction, or indeed by using a navigational query in order to locate a certain website.

Some queries may combine the types by referring simultaneously to the desired Goal and the website that the user wants to go to in order to find the information or buy the wares, as in the following examples.

8.38	**API wikipedia** {$\text{API}_{n.sg}$} {$\text{wikipedia}_{n.prop}$}	TDC (1)

[174] Situation facets were discussed in section 2.3.3.

8.39	**google wozniacie tennis** {google$_{n.prop}$} {(wozniacie$_{n.prop}$) < (tennis$_{n.sg}$)}	TDC (1)
8.40	**google billeder fra krigen** 'google images from the war' {google$_{n.prop}$} {billeder$_{n.pl}$ fra$_{prep}$ krigen$_{n.sg.def}$}	TDC (1)
8.41	**tdc play pink fucking perfect** {tdc play$_{n.prop}$} {[pink$_{n.prop}$] [fucking_perfect$_{n.prop}$]}	TDC (1)

Apart from the navigational references, ex. 8.38-8.39 refer to topics, while in ex. 8.40 the user is looking for media, and in ex. 8.41 the Goal is likely to be *BUYING/DOWNLOADING A PRODUCT*, in this case a track by the singer-songwriter Pink, from the music downloading service TDC Play.

As alluded to above, a major reason for the confusion of terms in standard query classification appears to be that definitions refer sometimes to what we have called *Goals* and sometimes to *Aims*. We shall now have a look at the formal difference between the two as proposed here. We shall define **Aims** in the following way.

> Aim: The desired response – such as the retrieval of a resource or the carrying out of an action – that the user intends the LAT to make for the user as a result of issuing a query. The Aim represents an attempt to satisfy the need associated with the Trigger behind the search.

An *Aim* is not the solution to an overall task, but rather the immediate result of employing the tool; it represents the desired effect of executing the search. Queries that refer to an action to be performed thus target the Aim. However, Aims may also be the resources themselves, such as references to documents, texts, lists or collections. This is not to be confused with reproductions of anticipated text, which do not target the Aim because they refer to the contents of the text rather than the resource in which it is couched. Instead, text anticipation targets what we shall call the **Object**. The Object is the actual resource in the shape of a document, a stretch of text, information treated within it, etc. as covered by the various resource facets. Reference to the Object is made by anticipation.

Object: The actual resource that the user desires to find and of which (part of) the contents are predicted, either because they are known in advance or because it is expected, hoped or hypothesized to have such contents. The Object is a manifestation of the Subject.

It is important to be aware of the difference between references to a kind of resource such as *price list*, *manual*, *novel* and *illustration* and references to the resource contents themselves; the former target Aims and the latter Objects. Naming the kind of resource amounts to requesting that such a resource be returned in response, while a reference to the contents is just that. The LAT cannot return contents; it can only return a resource containing the contents in question. Here is an example.

8.42	**prisliste** {$\text{prisliste}_{\text{n.sg}}$}	'price list'	TDC (349)

The point of this query is evidently to retrieve the mentioned resource, and this retrieval is an Aim in itself, but reference is made to the Object rather than the action aimed for. When this happens, Object and Aim all but merge. In Broderian terms, these are a kind of *navigational* queries. Obviously, such terms may sometimes be combined with other targets, as in the following examples where they occur with topical Subject-targeting references.

8.43	**prisliste mobiltelefoni** {$\text{prisliste}_{\text{n.sg}}$} {$\text{mobiltelefoni}_{\text{n.sg}}$}	'price list mobile telephony'	TDC (18)
8.44	**brugermanual mobil** {$\text{brugermanual}_{\text{n.sg}}$} {$\text{mobil}_{\text{n.sg}}$}	'user manual mobile (phone)'	TDC (113)
8.45	**forbindelsesdiagram multistik** **Diagram multistik** {$\text{forbindelsesdiagram}_{\text{n.sg}}$} {$\text{multistik}_{\text{n.sg}}$}	'wiring diagram connector' 'Diagram connector'	TDC (1) TDC (1)
8.46	**liste med landekode** {$\text{liste}_{\text{n.sg}}$ med_{prep} $\text{landekode}_{\text{n.sg}}$}	'list with country code'	TDC (3)

Goals shall be defined as follows.

> Goal: The state (possibly focusing on having accessed a resource, some information on a topic, or a desired object, etc.) that the user wishes to enter at the end of a search session as a result of having issued one or more queries and accessed one or more suggested resources (possibly complemented by some ensuing interaction with a website). The Goal represents a solution to the problem associated with the Source situation.

This carefully stated and somewhat complex definition of Goals covers many different scenarios. The user's Goal may be a state of having increased one's knowledge about a topic, or of having solved a problem, etc. It may even be the condition of having reached a certain website for the sake of experiencing it (cf. Rose & Levinson's (2004) *entertain* goal). In all of these cases, entering a state where satisfaction prevails is the true Goal, as opposed to the method of getting there, which is covered by the suggested notion of *Aim*.

Jansen *et al.* (2008) come very close to recognizing the distinction between Aims and Goals in their classification of navigational queries, which they note may lead the user to either an informational or a transactional site. In this case, the user's Aim is *NAVIGATION TO A SPECIFIC SITE*, while the overall Goal is either informational or transactional, "so, one can view navigational searching as an expression of an intermediate intent aimed at satisfying some larger searching goal" (Jansen, Booth, & Spink, 2008, p. 1260). They even use the terms "aimed" and "goal" in a manner that conforms to the terminology introduced here; navigational search is considered to involve Aims in the suggested framework, with the other Broderian types representing potential Goals following such navigation. The present author thinks it highly improbable that navigation to a site could ever be the Goal of a search in itself in the sense of *entering a state where satisfaction prevails*. Rather, there must be some point of going there, even if the purpose is simply to experience the site in order to make sure that it exists, delight in its visual appearance or check how the webmaster in question tackled the challenge of constructing the site. The Goal in these cases will be of a psychological kind (as indeed most Goals are, satisfaction being a psychological phenomenon), while the Aim is certainly to be taken to the site in question and nothing beyond that.

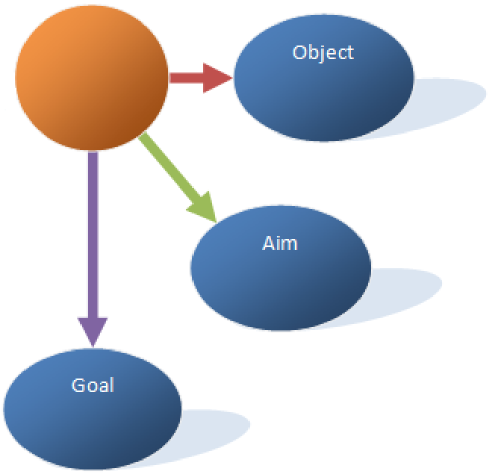


Fig. 8.2. Partial target model. The *Object* represents the actual contents that can be anticipated, the *Aim* is the outcome of the search process itself, and the *Goal* is the final state where the problem that triggered the search should be resolved. Query terms may refer to either of these targets. The model is expanded below.

As is evident from some of the more problematic examples discussed, performing a proper Target Analysis is not easy; one target can easily be mistaken for another by the untrained analyst. Two kinds of distinctions in particular may deceive the analyst; one is represented by the difference between Trigger and Aim (the psychological idea of what you need vs. a preconception of what will actually satisfy that need), and the other by that between Object and Aim (the resource sought for vs. the action of retrieving it). It is fully possible for several targets to be psychologically present simultaneously (fig. 8.2). For instance, users may employ anticipation of the *Object* text with the *Aim* of navigating to a site with a particular *Goal* in mind. On the other hand, the query may target a Goal with no particular mention of how to reach it; indeed, the user may have no idea about how to attain the Goal and may be hoping to receive information that will clarify this. When it comes to the expression of queries, only one target may normally be chosen for every term. The exception is Trigger and Aim which in contrast to the other targets are often signalled by form rather than content. This makes it possible to combine one of these form-borne targets with one that is signalled through semantic content. For example, it is possible to target the Object of the search by referring to the actual document or its contents, but at the same time the Aim is secondarily signalled by this very use of a reference to a feature of the Object; the user clearly wants the system to carry out the action of retrieving and displaying the resource in question. Similarly, using a topic label indicates what topic the user wants information on, thus targeting the Subject, but the very use of such a label also secondarily signals the Trigger; the user obviously wants information on a topic rather than a fact, a straight answer to a question or a specific resource. This situation exactly parallels the one involving Aim and Goal. On the other hand, it is not possible to signal both Source and Subject with the same term, for instance, since both use the semantics of the term as their expressive medium. A topical search will be said here to target the Subject, even though it effectively signals the Trigger as well. Similarly, an anticipating search will be said to target the Object although it also signals the Aim. This is because the semantics of the term in question code the actual topic or resource targeted and hence the Subject and Object interpretations provide more information in these cases.

The proposed distinction between Goals and Aims is somewhat similar to what Xie (2007) calls *work task* or *leading search goal* and *search task* or *current search goal*, respectively[175]. The crucial difference is that she considers these to be essentially of the same type, so that the current search goal may become a work task with a subtask which acts as a new search goal. This confounds the basic difference in kind between search goals and work tasks; most tasks – such as *WRITING A THESIS* – can never act as search goals since no search result can lead to a fulfilment of the Goal. On the other hand, writing a thesis may be the ultimate **Objective** – a task that the user wants to be able to carry out after having successfully reached the Goal and thus completed the search process in full. A Goal in this instance might be to understand the topic clearly and so to be ready and able to start or continue writing the thesis. The difference is one between actor and instrument; a hammer can never be made to build a vessel, but it can fasten the nails.

Objective: The thing that the user desires to be able to do after having performed a successful search. The Objective satisfies the emotional need associated with the Motivation behind the search.

Thus, while the Goal in our terminology is what the user intends to achieve with the help of the IR tool, the Aim is what he or she wants the tool to carry out. The Objective represents what the user wants to be able to do as a consequence of having performed a successful search.

It is common for different levels of tasks to be alluded to in discussions of user goals. Obviously, a search is indeed part of a larger work task that instigated it by triggering a need. Such a work task may again be part of an even larger task or project, which again may contribute to an ultimate task such as earning your living and leading a satisfactory life. There is no limit to how many levels can be declared in such a taxonomy. To simplify matters, we have said in Part I that there are basically two **spheres** at work in search – the **internal user context** in which the search is conducted, and the **external user context** in which the actual work is taking place. The two spheres are bound together by the fact that the Goal is the same in both; there is no need to search for something different from what you think you need in the external situation, and consequently the internal Goal is inherited from the external sphere. On the other hand, the **internal Aim** is to generate a response of a certain type from the LAT, while the **external Aim** is to deploy this response in performing the task or solving the problem. We

[175] See also Ingwersen and Järvelin (2005a).

shall mostly be talking of Goals and Aims without making a distinction between the internal and the external sphere, as they are so tightly bound together by the common Goal anyway, but it is important to be aware of the differences, especially because most authors make this distinction rather than the one between Goals and Aims which is here suggested to be the more fundamental one when dealing with query formulation (but not necessarily as regards other behavioural characteristics of searchers).

Whether users will be most helped by attempting to fulfil their Goals rather than helping them attain what they are aiming for is a vexed question. In many situations, this is clearly the case. For instance, users who want to buy a certain item and consequently search for a web shop from which it might be obtainable would clearly be helped by being pointed to an off-web shop that stocks the item in question, in the event that it cannot be bought from any known web shop[176]. Receiving the address and opening hours of such a shop would be far more helpful than no result at all. On the other hand, users may be annoyed at not being obeyed to the word, especially if the system is not perfect, which is after all a most likely scenario (fig. 8.3).

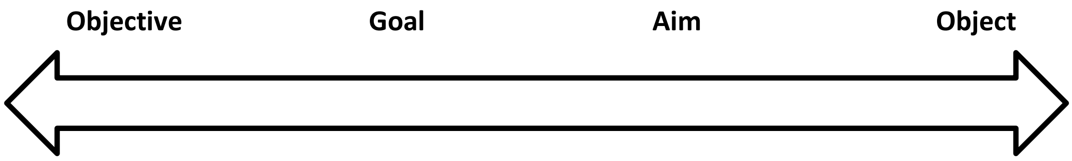


Fig. 8.3. A LAT may be tuned to listen to what users ask for precisely or attempt to help them obtain what they actually need. The system designer must choose a point on this scale to aim for. Traditional search engines are only able to help identifying the Object (note that the Object is faceted, and consequently this includes what might be termed *anticipated topics* in addition to anticipated text). Some are more advanced and attempt to retrieve documents that the user is thought to prefer, i.e. they are designed to support Aims. It is possible to go further and attempt to help users reach their Goals and become satisfied, or even to try to understand the ultimate Objectives of the search, but this incurs an increased risk of giving the users the wrong thing while ignoring what they actually ask for, so it is important to restrict such attempts to particularly clear cases, or alternatively to present these suggestions in a manner different from the standard results so that the user is free to choose.

We have now introduced all of the eight possible targets covered by the suggested framework, and it is time for a complete overview to be given. The eight potential targets are as follows (they only become actual targets the moment a query term refers to that particular aspect of the need).

[176] Note that Broder (2002) assumes that an intention of shopping off the web results in informational queries. This may not be entirely true, since if the shop is known in advance the user will be likely to perform a navigational query in order to find facts on the shop, such as the address or opening hours. Also, even when shopping from a web shop informational queries are likely to be used in many cases in order to find information on available products.

- **Motivation**: The affective drive towards a possible personal gain that triggers the user to approach a LAT in the first place.
- **Source**: The problematic situation or sense of lack that makes the user realize that a search needs to be made.
- **Trigger**: The actual need that triggers the user to search for something particular.
- **Subject**: The topic that the user chooses to search for in order to reach desired resources.
- **Object**: The actual ideal resources that the user has in mind.
- **Aim**: The ideal problem-solving outcome of the interaction with the (extended) system that the user envisages.
- **Goal**: The ideal, problem-free state that the user desires to reach through interacting with the (extended) system.
- **Objective**: The ultimate purpose of searching characterized by personal gain or new abilities due to the (imagined) successful search.

Each potential target is explained and discussed below. They all have an abbreviation used in the formal representation of queries in the expanded version used henceforth (fig. 8.4).

C1	Stage 1 cause: Motivation
S1	Stage 1 state: Source
N1	Stage 1 need: Trigger
T1	Stage 1 topic: Subject
T2	Stage 2 topic: Object
N2	Stage 2 need: Aim
S2	Stage 2 state: Goal
C2	Stage 2 cause: Objective

Fig. 8.4. Abbreviations used for symbolizing targets in the formal notation of queries.

Together, they make up a cohesive model of the search process from the first vague motivations to the ultimate objectives, a model that explains and illustrates why queries are so diverse. Before going more into this framework, however, we need to look into the referential processes that it rests upon.

8.1.4 Modal and aspectual reference in use

In natural language, Stage 1 can be equated with the indicative and Stage 2 with the optative. Some kinds of Stage 2 queries are directly comparable to imperatives, and indeed imperative morphology is very common in this type of queries. European languages tend to confine mood, aspect and tense to verbs, but in some other languages modal, aspectual and temporal meanings can be assigned to nouns in the same way, as in the following sentences from the North American language Makah where the word for ‘dog’ acquires a conditional form when representing a need (Mithun, 1999).

8.47	q̓idi·ƛ̓ʾu dog-PAST-INDIC.3	’It was a dog’
8.48	ʾusubas q̓idi·ƛqey need-INDIC.1sg dog.COND.3	‘I need a dog’

Queries, too, allow this non-specific use of nouns, though no explicit marking is used in English or Danish. In L^{WID}, for example, many users search for the jobs they would like to get, or the work tasks they would not mind performing in future, and this may be considered an instance of Stage 2 searching. The 10 most commonly occurring references to jobs and desired tasks in L^{WID} are shown in fig. 8.5. These are not topical queries, as the users are not looking for information on drivers and cleaning etc., but rather for material that will help them reach a stage characterized by a manifestation of these concepts. Consequently, they are instances of Stage 2 query reference.

Query	Subtype	Frequency
driver	S2: job description	121 x 10
cleaning	S2: work task	116 x 10
teacher	S2: job description	89 x 10
nurse	S2: job description	85 x 10
architect	S2: job description	83 x 10
chemist	S2: job description	57 x 10
designer	S2: job description	53 x 10
cleaner	S2: job description	49 x 10
accountant	S2: job description	48 x 10
welder	S2: job description	48 x 10

Fig. 8.5. The 10 most commonly occurring references to desired jobs and work tasks in L^{WID}. The total number of such queries in L^{WID} was 4296 x 10, distributed over 1461 distinct strings in the reduced set. The dataset was reduced to 10%, which is why frequencies are given in this manner.

In addition to the modal categories, there are aspectual distinctions within S1 and S2, as each can be regarded from either point of view. This, too, is particularly clear in the analysis of the L^{WID} log, because when people search for job postings they may either describe themselves and their situation (S1 as seen from S1) or the way they imagine that a potential employer would see them, as possessing certain titles, qualities and competencies (S1 as seen from S2). Alternatively, they may describe what they currently need (S2 as seen from S1), or search directly for the job they want, as something that is independent of their needs (S2 as seen from S2). These expressive methods can be likened to aspectual differences in language such as the use of retrospective and prospective aspect in English as shown above. After all, it is exactly when referring to this kind of differences that we normally need aspectual morphosyntax in language. Compare for instance the expressions in fig. 8.6 with the examples from L^{WID} that follow.

	Stage 1 (Indicative)		**Stage 2 (Optative)**	
	Declarative	**Retrospective**	**Prospective**	**Desiderative**
S1	**I am Spanish**	(I have been being Spanish)	(I am going to need to be Spanish)	(I want to become Spanish)
S1’	(I am studying biology and working with java)	**I have been studying biology and working with java**	(I am going to need to work in biology with java)	(I want to work in biology with java)
S2’	(I am having a student job in Copenhagen)	(I have been having a student job in Copenhagen)	**I am going to need a student job in Copenhagen**	(I want to get a student job in Copenhagen)
S2	(I am a driver)	(I have been being a driver)	(I am going to need to be a driver)	**I want to become a driver**

Fig. 8.6. Comparison of the modal and aspectual categories found in language and in search. Most of the expressions are perfectly acceptable sentences, but not all situations portrayed by them are likely to occur in the context of looking for a job.

⬤ S1 from S1 (Declarative)

Users looking for a job may describe themselves or the (often problematic) situation in which they find themselves.

8.49	**unemployed** {S1:Situation: unemployed$_{n.sg}$}	WID (9)

Given that the Goal of finding a job is inherent in the use of the tool, these descriptions do not necessarily represent the problem that needs solving but often simply informs the system of their background against which a job posting can be matched.

8.50	**spanish** {S1:Origin:Nationality: spanish$_{n.sg}$}	WID (2419)
8.51	**young** {S1:InitialState: young$_{n.sg}$}	WID (1)

Perhaps surprisingly, this category actually accounts for a range of the very commonest query strings in L^{WID}, especially in the form of nationalities (fig. 8.7).

	Subtype	Frequency
spanish	S1: Origin: Nationality	242 x 10
engineer	S1': Field: Title	158 x 10
french	S1: Origin: Nationality	141 x 10
russian	S1: Origin: Nationality	138 x 10
farm	S2: Environment: Workplace	130 x 10
english	S1: Origin: Nationality	127 x 10
driver	S2: Goal: Post	121 x 10
cleaning	S2: Goal: Task	116 x 10

Fig. 8.7: The eight most commonly occurring queries in L^{WID}. Many nationality-designating terms, including those shown here, are ambiguous between terms for nationalities and linguistic competencies. Thus, *spanish* could be a reference to a linguistic skill, or even a subject that the user was contemplating to teach in Denmark. However, it is unlikely that all these people are looking for jobs in education since the number of explicit references to teaching is insignificant. Rather, most of them probably enter their nationality in the hope of finding something suited to people from that country, being unable to foresee what jobs they might possibly be able to get in a foreign country. This is clearly the case whenever queries contain the names of countries and cities, but nationalities remain ambiguous and are therefore treated here as a separate inherently ambiguous subtype. The total number of S1-targeting queries in L^{WID} is 1758 x 10, distributed over 279 distinct queries in the reduced dataset.

Other less common subtypes present in L^{WID} include the *initial state* (the user's class of person) and *situation* (problematic state), as shown. These are not always easy to distinguish, and they are relatively rare. Some additional examples are given in Appendix C along with samples of the other types that occur in L^{WID}.

● S1 from S2 (Retrospective)

Another possibility is for users to refer to themselves as they imagine that a potential future employer would see them, i.e. as possessing certain qualities and competencies, or lack thereof. This retrospective aspect will be symbolized by S1' in the formalized notation.

8.52	**no experience** {S1':Experience: no experience$_{n.sg}$}	WID (21)
8.53	**biology** {S1':Field:Academic: biology$_{n.sg}$}	WID (347)
8.54	**java** {S1':Skills:Tool: java$_{n.prop}$}	WID (913)

In the analysis of L^{WID}, retrospective subtypes employed include previous *experience* (amount of experience, certificates earned), competencies and *skills* (subjects on which the user has knowledge, tools that the user knows how to use, linguistic skills, subjects that the user would be able to teach), *field* (titles acquired, academic fields in which the user was educated), and *school of thought* that the user belongs to.

● S2 from S1 (Prospective)

Yet another way of performing the request is to refer to what you need, given your present condition. This prospective aspect shall be symbolized by S2'.

8.55	**student job** {S2':Need:NeedType: student job$_{n.sg}$} / {S2':Need:NeedType: (student$_{n.sg}$) (job$_{n.sg}$)}	WID (242)

8.56	**seasonal work** {S2':Need:NeedType: seasonal work$_{n.sg}$}	WID (102)
8.57	**Copenhagen** {S2':Locality: Copenhagen$_{n.prop}$}	WID (123)

The subtypes met with were *locality* (desired geographical workplace), *needs* (type of need, level of seniority), *conditions* (desired employment conditions, working hours etc.), and *method* (methods, systems or tools that the user wants to work with).

⬤ S2 from S2 (Desiderative)

Finally, users may simply refer to a desired job as something that exists independently of their needs.

8.58	**driver** {S2:Goal:JobTitle: driver$_{n.sg}$}	WID (1205)
8.59	**cleaning** {S2:Goal:Task: cleaning$_{v.nom}$}	WID (1156)
8.60	**agriculture** {S2:Sector:Area: agriculture$_{n.sg}$}	WID (588)
8.61	**farm** {S2:Workplace:Environment: farm$_{n.sg}$}	WID (1298)
8.62	**pig** {S2:Core:Material: pig$_{n.sg}$}	WID (267)

This aspectual type is the most commonly occurring in L^{WID} overall, when all queries in the analysed log are included; they are simply more diverse than the references to nationalities, which are consequently found among the most frequent individual query strings. Subtypes in L^{WID} include *workplace* (desired workplace/environment, company or type of enterprise), *goal* (job title, work task), *core* (material, focus), and *sector* (sector, department or activity occurring across sectors).

Furthermore, the subtypes of *earned degree* and *training position* were assigned to an inherently ambiguous type that belongs to both S2 and S2' because a term such as *PhD* may refer either to a degree that one has achieved or a job position that one wants to get, and a similar argument can be made regarding the various kinds of assistants mentioned in many L^{WID} queries.

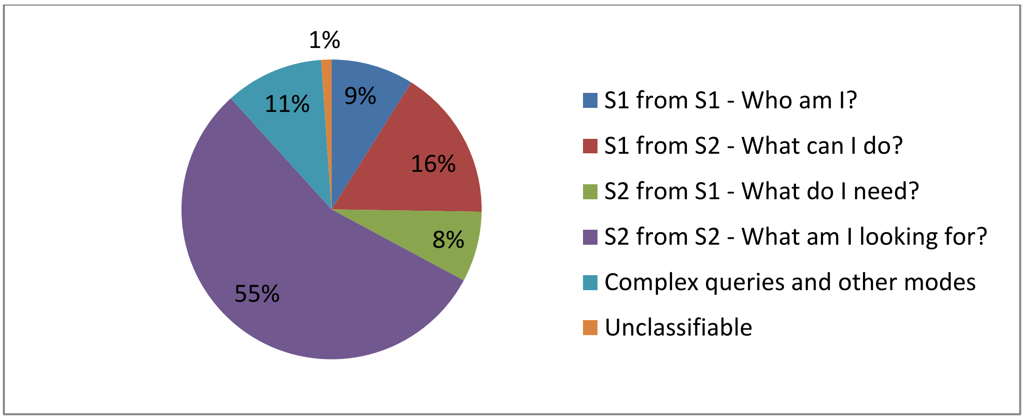


Fig. 8.8. The distribution of stages in L^{WID}. The chart is based on 10% of all separate searches performed during one year, using a method of reduction set out in section 3.1.2. The category *complex queries* mostly covers mixing types which include several different references in one query, but also the inherently ambiguous types and a few that reflect a wish to receive information on a topic or retrieve a certain resource. These are very rare in this particular log material due to the specialized domain that it represents. The unclassifiable queries were simply too linguistically ambiguous or unclear to be classified.

In L^{WID}, the various moods and aspects turn out to have the distribution shown in fig. 8.8. A more detailed account of this study is given in section 10.1. Clearly, people do not all search in similar ways; rather, even in such a narrow domain as that represented by the Workindenmark site, there is an astounding variability in the way queries are constructed; there is clearly not just one way of searching for a job. The retrospective S1'-queries ostensibly refer to the user's current state, but they really describe the goals (here, the jobs) through the user. Similarly, the prospective S2'-queries appear to describe the goals (e.g. the geographical location of a desired job), but they really portray the users through their goals (the user wants a job not too far from home). Consequently, aspectual distinctions are dependent on the domain of the search, because the fact that non-retrospective S1-queries simply describe the initial state with no reference to the problem that needs solving means that such queries are only likely to occur in a context where it is entirely clear what the problem is, such as a job portal. In most environments, then, all S1-queries are retrospective. The difference between the S2 aspects are often harder to establish, but one example is the difference between searching for *billig mobil* 'cheap mobile phone' and referring to a particular model of mobile phone.

8.63	**billig mobil** **billig mobiltelefon** **billig mobil telefon** {S2': billig mobiltelefon$_{n.sg}$}	'cheap mobile phone'	TDC (13) TDC (5) TDC (3)

Both aspects of Stage 2 are thus frequently present side by side.

8.2 The Stage-based Intent Model

Classifying intent according to its intrinsic qualities, such as what the user's intentions were, or what kind of situation triggered the search, is almost impossible, due to the immense difficulties of drawing a line between different intentions or situations. We have therefore followed an entirely different approach by devising a model of the search process that encompasses causes and goals, topics and resources, and classifying queries according to what *category* among these they refer to, rather than the actual descriptions of problematic situations and solutions etc. given in queries. Some refer to the Source of the search (the situation that prompted it) and others to the *Subject* (the topic on which information is desired). Others again refer to the user's *Aims* or *Goals*, etc. This provides us with a finite number of query types. The category chosen for query reference we have called the *target* of the search. The model as a whole is referred to as the **Stage-based Intent Model**.

8.2.1 The individual targets

Above, the various potential targets were introduced in a somewhat unsystematic manner for pedagogical reasons. Below, they are discussed one at a time, and their role in the model as a whole is made clear. The strength of the model is that it mirrors the search process from the first vague Motivation over planning and action to the possible attainment of Goals and perhaps even realization of the ultimate Objectives, while at the same time making clear exactly what stages in the process users may target with their queries, thus unifying behaviour and expression in a unique way. The process proceeds as follows.

● **Motivation**

The potential target that we have called *Motivation* covers the emotional state and mental needs that caused the user to consult the LAT in the first place. It represents a very general urge to be informed, advised or entertained, or in some cases to acquire profits, status or power.

> In fact, the concept of *motive* may be of general use in the study of information-seeking behaviour since, if we assume that, for whatever reason, a person experiences an *information need*, there must be an attendant *motive* actually to engage in such behaviour. (Wilson T. , 1997, p. 553; emphasis original)

There can be many such kinds of Motivation in our lives; among them, Morgan and King (1971) suggest the following types: 1) *physiological motives* such as hunger and thirst, 2) *unlearned motives* such as curiosity and sensory stimulation, and 3) *social motives*, such as the desire for affiliation, approval or status. These seem to be very much akin to what we mean by the term *Motivation*. Mostly, Motivations are just a kind of background noise, but occasionally they build up into a distinct urge that causes us to act. Typically, this change happens at a so-called *breakdown* in a problem-solving process[177]. This term is used in a technical sense in reference to "a disruption in the normal functioning of things forcing the individual to adopt a more reflective or deliberative stance toward ongoing activity" (Koschmann, Kuutti, & Hickman, 1998, p. 26). Performing some kind of information retrieval is a normal thing to do in such circumstances, whether this involves an IR tool or not.

For an example of Motivation as a stage in the formation of a need, consider the following Subject-targeting query.

8.64	**låneansøgning** {T1: låneansøgning$_{\text{v.nom}}$}	'loan application'	TDC (2)

This search was clearly motivated by economic issues, and so Motivation in the sense established here plays an important part, although it was not chosen as the actual referential *target*. Motivational targets are in fact very rare, as people do not usually employ phrases such as *entertain me* or *I need some advice* in queries. After all, it would be felt that some more explicit need, problem description or topic must be provided, as indeed is the case in ex. 8.64, where the user chooses to target the Subject and thus perform a topical search.

At times, Motivation may stem from other people asking or expecting the user to perform, such as teachers urging their students to read, or to buy books. Motivation is thus the entry

[177] It is possible to claim that this is so by definition if one will accept that, say, surrendering to a craving for entertainment involves a breakdown. This seems indeed to be the position taken by Ingwersen (1992, p. 46), who states that "to users the breakdown situation initiating IR takes place in their 'conceptual knowledge', implying that they are unable to solve problems or *fulfill emotional or interest-related goals*" (emphasis added).

point into search for external influence such as biological and social urges, but also emotional effects. People are only occasionally conscious of their Motivations in the sense of the word proposed here, but Motivation affects what they search for in a fundamental way. To what extent it has an effect on how they express their needs is less clear, though it seems that people do occasionally target their Motivation, if the following are to be considered cases of Motivational targets.

8.65	**jeg er utilfreds**	'I am dissatisfied'	TDC (1)
	utilfreds	'dissatisfied'	TDC (1)
	utilfreds kunde	'dissatisfied customer'	TDC (2)
	{C1: utilfreds kunde$_{n.sg}$}		

Whether this is to be interpreted in such a way depends on whether being dissatisfied is considered a need-generating situation or merely an emotional impetus. Something cannot be a Motivation if it depends on a situation where a particular need arose; a need generated by a situation characterized by a lack is the Source of the search rather than its Motivation. True, a need for entertainment is a kind of need and might hence be thought to belong with the Source, but it is one that is presumably always present to some extent. Sometimes the urge is stronger and we become conscious of it, but it is independent of any concrete information need, and therefore it is a Motivation rather than a Source situation. The same is true of a weak version of the need for advice. In a sense, we are always in need of advice, in general terms, so many searches are advice-motivated. Specific situations may arise in which we need more particular advice, and when this happens, it is not merely a case of Motivation, but of problematic Source situations and the advice *Trigger*. Finally, as an eager researcher I would like to believe that all people invariably have some general craving for information, which is likewise an ever-present Motivation to search, though it is also obvious that this Motivation only occasionally develops into a more concrete need deriving from a Source situation.

> That needs may have a cognitive component (as distinct from, for example, physiological needs such as hunger and thirst) is recognized in the concept of the *need for cognition*: the need to find order and meaning in the environment, which is also expressed as the need to know, curiosity, the desire to be informed. (Wilson T. , 1997, p. 553)

When a particular need appears in the form of a Source situation, the urge to be informed is instantaneously strengthened and concretized.

⬤ Source

The Source is much more concrete than the affective Motivation. It basically consists of a situation characterized by a problem or sense of lack, sometimes rather experienced as a desire. This has also been characterized with a much-cited sentence originating with Mackay (1960) as "a certain incompleteness in [the user's] picture of the world – an inadequacy in what we might call his 'state of readiness' to interact purposefully with the world around him". Users tend to be much more aware of the Source than of the Motivation, which is highly emotional. Some authors call the Source the *problem space*. Ingwersen (1992, p. 113) characterizes it (or rather the stage in the process at which it is acknowledged) as "a situation specific state of mind in which the individual recognizes its lack of knowledge, e.g. in order to choose between possibilities of action, of solutions to problems, or in relation to the fulfilment of factual or emotional goals". Some of these Goals are fed directly from Motivation (e.g. the need for entertainment), while others arise in the situation as such.

The situation that affects the user is not the actual physical occurrence, but rather a mental reconstruction of it made by the user, and so it, too, is psychological, though not so much affective as rational, or perhaps rather sensorial. However, in spite of its rational character, a Source is not something that the user consciously constructs after considering the possibilities offered by search. Rather, it is something that presents itself as the initial state, the starting point of the search, Stage 1.

> [T]here is not at first a situation and a problem, much less just a problem and no situation. There is a troubled, perplexed, trying situation, where the difficulty is, as it were, spread throughout the entire situation, infecting it as a whole. (Dewey, 1933)

The Source may be a materialized or manifest version of a Motivation, and as such some situations imply a need for information, others a need for advice, and some perhaps a desire to be entertained (just think of a situation characterized by boredom), and so on. It follows that boredom is *caused by* an initial need for entertainment, rather than the other way round.

⬤ Trigger

Before initiating the actual search process the user must settle on a need that can feasibly be satisfied by such means. At this point, therefore, the user effectively defines his or her needs more or less consciously, for the purposes of the search. It may well be that what the user really needs is something entirely different, but that is not part of the search process; that belongs with the relevance assessment after results have been seen and evaluated.

From this point on in the process of need development, users decide what kind of help is needed. This is also the point where they decide whether to simply describe the problematic state and hope for a solution to be presented in the search results or to go for a particular solution. Here, they choose between searching for a fact, asking a question, or just throwing in a reference to a topic and hoping for some useful information to be returned. The choice may also be between searching for a resource to experience for the fun of it and searching for interesting and entertaining information, or between targeting the acquisition of a product and merely locating information on how to obtain it. As such, it is a crucial moment in the search process. The Trigger is thus neither a situation nor an actual topic but a *type or way of looking for a solution* that represents the need as conceived by the user. The Trigger can be the answering of a question, information on a topic, locating a fact, retrieving a resource, etc. The various types are explored in section 8.3. The way Triggers are targeted is by using certain morphosyntactic forms, such as questions, imperatives, topic labels, names of resources, etc.

● Subject

When the user has decided on a method, a Subject may present itself in the form of a topic on which to search for information. This is not always the case; frequently advice is acquired without reference to particular subjects, and entertainment may likewise be gained without a topic in mind, and so on. Even so, targeting a Subject is a very common scenario. The Subject as a potential target represents the topics or interests which are at the centre of the need, i.e. the *intrinsic information need* (Ingwersen, 1996); it does not contain situations but only labels. As noted by Ingwersen, the intrinsic information need may be quite stable, even when other aspects of the user's understanding are in flow as discussed by Kuhlthau (1993).

Use of Subject-targeting is augmented by the *label effect* which incites people to express themselves in the form of topic labels even when their need cannot easily be converted into topics.

● Object

While the Subject is the internal representation of the needed information, the Object is its external manifestation. The Object consists of the actual resources that the user is looking for. Nonetheless, it is a psychological entity, because these resources are ideal ones in the mind of the user, given that the search has not been performed yet and consequently no results have been returned. Most frequently, it involves strings that the user wants or expects to find or refind in documents, but it can also be descriptions of media resources or even physical objects that the user wants to acquire, or references to their parts.

Object-targeting is the most primitive way of searching, but it is also the technique that leaves the user maximally in control of the results.

● Aim

The Aim stands for the kind of response that the user hopes to achieve by searching. It does not in itself represent the final outcome of the search but the change itself, such as returning useful information, sending a desired product or performing a requested action. This is of course closely related to the satisfactory state resulting from it, but it is not identical to it. While the Trigger represents the user's idea of what is needed, the Aim is what the user thinks should ideally result from the search in order that this need may be satisfied. Thus, the Trigger is the internal representation of the need, whereas the Aim is its external manifestation.

One type of Aim-targeting queries refer to (parts of) the actual resource that should be retrieved. This indicates a wish to have this resource retrieved. Only meta-references to the resource belong here, as opposed to references to the contents of such passages, which target the Object.

● Goal

The Goal is the ideal state resulting from a search – one that is characterized by satisfaction. It is what the user wants to achieve by searching and applying the results. A Goal does not necessarily have to be obtainable simply by searching; it may require further application of the results. However, it is the natural endpoint of the process of which the search is a crucial part. While the Source is the initial state characterized by a problem, the Goal is the ideal replacement for this situation, a state where the problem is resolved. Hence, the Source is the internal representation of a problem, while the Goal is the external manifestation of its solution.

● Objective

As opposed to Goals, Objectives are not restricted by the capabilities of the tool; they represent the ultimate purpose of the search in terms of real-world achievements, typically in the way of being able to carry out some subsequent task. The psychological sense of ability is thus at the heart of Objectives; it is a state in which the feeling of being restricted in some way that may have characterized the Motivation is resolved. In some cases, it is not so much a question of abilities as personal gains of other kinds, such as entertainment, knowledge and understanding, profits, status, or power; all the things that drive people to act. It thus represents an ideal endpoint of the Motivation that started the process, and you might again say that while the Motivation is an internal representation of the reasons for searching, the Objective is their external manifestation. Targeting the Objective appears to be rare, but for natural reasons Objectives figure prominently in people's conceptions of their needs. Like Motivations, they represent the outmost limit of what a query may refer to, only at the opposite end.

For simplicity, the above discussions have been concerned with one target at a time. This may have given the impression that there is never more than one target in the same query string, which is not true. Most queries do only have one target (except for the co-occurence of Trigger or Aim with another target on the same term), but a few combine several targets, such as one term targeting the Subject and another targeting the Object, or one targeting the Aim and the other targeting the Subject as in the following example.

8.66	**Kontakt TDC driftsproblem** 'Contact TDC operational problem' {N2: Kontakt$_{v.imp}$ TDC$_{n.prop}$} {T1: driftsproblem$_{n.sg}$}	TDC (1)

8.2.2 The dimensions of a need

The eight possible targets are not random phenomena but strictly ordered elements in a comprehensive model. It is clear from the above descriptions that the different kinds of intent appear in pairs of restriction and resolution, or need and gratification, as it were. The targets are thus both part of a progressive timeline from the first Motivation to the ultimate Objective and occur pairwise in a pattern of need and resolution (fig. 8.9).

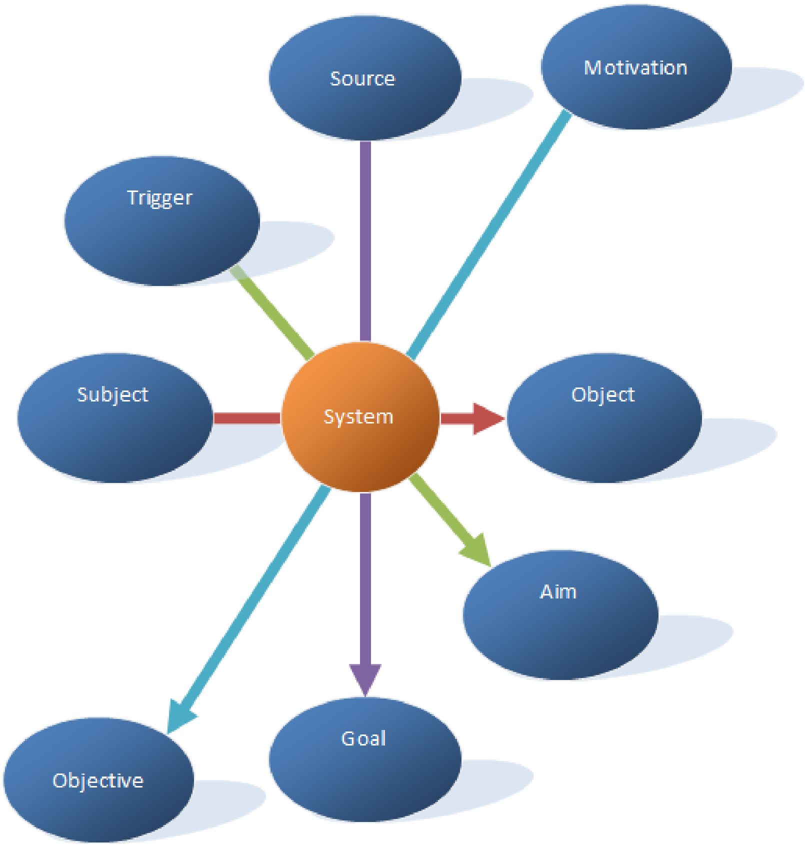


Fig. 8.9. A general illustration of the suggested model of the search process. Motivation, Source, Trigger and Subject belong to Stage 1, while Object, Aim, Goal and Objective are the Stage 2 counterparts of these targets.

The Stage-based Intent Model accordingly incorporates four dimensions along which the user moves during a search. One is from the Motivation that fuelled the searching activity in the first place (the *Motivation*) via system interaction to the ideal final sense of ability and resolved restrictions (the *Objective*); the second leads from the situation that prompted the search (the Source) via system interaction to the ideal resultant state where the task has been completed and satisfaction achieved (the *Goal*); the third is the process from the more or less conscious need for help (the *Trigger*) via system interaction to the ideal response as imagined by the user (the *Aim*); and the fourth and final dimension stretches from the domain of relevant topics (the *Subject*) via system interaction to the ideal resource that will elucidate them further, including information that can be drawn from it (the *Object*). The path from Subject to Object (from topic to resources) is the shortest and most direct and unproblematic one, and the one from Motivation to Objective is the longest, spanning the entire process from the affective state that made the user search in the first place to the Objective that was the ultimate purpose of the search, as illustrated by the length of the arrows in fig. 8.9. Resources may be tuned towards different dimensions in the Stage-based Intent Model. Fiction targets the Motivation-Objective dimension, whereas factual resources exist to satisfy needs in other dimensions. There are many subtypes of such resources. Some target the Source-Goal dimension by attempting to help users in certain situations, some attempt to answer people's questions or guide them through the performance of onsite actions (Trigger – Aim), and some deal with specific subjects and general pieces of information such as price lists (Subject – Object). In principle, such differences could and should be provided as metadata available to the LAT.

If we compare the Stage-based Intent Model to the stages of search behaviour discussed by Kuhlthau (1993) (section 2.1.3), we see that her *initiation* stage corresponds to the recognition of what we have termed the Source, while the *selection* stage corresponds to the identification of a Trigger and Subject, and *exploration* involves the Subject and Object. Note that the two models are complementary in that Kuhlthau's terms refer to stages of the search process whereas our terminology deals with the actual expression that the user ends up using. It is possible to correlate them because the expression may target the different clues presenting themselves to the user during the process of finding out what is needed in a given situation, but the Stage-based Intent Model presented here does not sketch the search process directly; only insofar as it is reflected in query formulation behaviour.

The *Subject* is a topic in the widest sense, i.e. a domain of interconnected ideas in the mind of a user. Accordingly, it incorporates the user's initial understanding of the problem and interests in various sub-topics. Part of the solution often lies in achieving a fuller understanding of the problem, and consequently the Subject-Object dimension is a path from initial understanding to the ideal resource that would provide the full necessary understanding of the topics identified as relevant by the user. Since the Subject is defined by the user and the Object is in fact a corresponding idealized entity created in the user's mind, other topics may in the end turn out to be much more informative, but that is a matter for relevance assessment and not part of the actual search process in a narrow sense[178].

It is important to understand that both Stage 1 and Stage 2 take place in the user's mind. The Source situation, for example, is not necessarily the real state of affairs, and hence includes affective characteristics which are not really part of the external situation, and the Goals and Aims are likewise mere suggestions presenting themselves in the user's mind. By definition they are ideal (hypothetical); achieving them may well turn out to be useless because of being based on mistaken premises, but that aspect once more belongs with relevance judgment and not inside the model of the search process as understood here for the purposes of log analysis. Note also that, because the eight endpoints are ideal

[178] Obviously, this depends on the definition of *search process* chosen. For the present purposes, it only includes that which the user may refer to through the query, or in other words, what can be the *target* of the search. For obvious reasons, this does not include relevance assessment, which takes place after results have been delivered. There is, however, nothing wrong with including relevance assessment in the definition of the term.

and not actual ones, the dimensions in themselves are not timelines, although they do each simulate the search process in their own way.

As mentioned earlier, Stage 1 reference is presumably restricted to problematic situations and clear cases of lack or instability. If such a situation is described, it is clear what the user wants, viz. to reach a more satisfactory Stage 2. If, on the other hand, Stage 1 is unproblematic in itself but the user wants to reach Stage 2 for other reasons, describing this desired Stage 2 is much more informative, because it may not be at all clear from Stage 1 what the desired Goal might be. There are thus relatively clear differences in intention and need type between Stage 1 and Stage 2 queries. Unfortunately, it is not always clear from the form of a query what stage it represents – cf. *iphone no sound*, which may refer either to the problematic Stage 1 or to the desired Stage 2, depending on whether the problem is turning the sound on or off.

Because both stages are mental phenomena, Stage 2 represents user intents just as much as Stage 1, and this goes for needs as well; what you need is neither restricted to desired resources, nor synonymous with the sense of lack you experience, but a combination of all of the eight potential targets. Authors tend to focus on much narrower definitions of the concept of user needs, and this keeps the resulting theories and frameworks from being applicable in a large proportion of searches. Often, needs and goals are mixed up.

> We adopt our basic definition of user needs and goals from (Jones & Klinkner, 2008), which defines a search goal as, 'an atomic information need, resulting in one or more queries'. (Kofler, Larson, & Hanjalic, 2011)

> [The] subjective character of need is evident, for example, in a definition by Burnkrant (1976) which proposes that need is "a cognitive representation of a future goal that is desired". (Wilson T. , 1997)

Needs and goals certainly are related, as clearly demonstrated by the Stage-based Intent Model, but they are hardly synonymous terms.

8.2.3 Pseudo-merger and the natural order of targets

Certain kinds of search seem to contain fewer possible targets than others because some of the points in the model are rendered all but equivalent. For instance, if the Aim of a search is to access a certain resource or acquire an object, then the Aim and the Object are very closely related. However, the Aim remains a desired change and the Object an entity. Users may then refer to either the Object itself or the action of acquiring (buying, downloading, etc.) it. Similarly, Aims may be as good as equivalent to the Goals if there is no further Goal with the search apart from performing the aimed-for action. Of course, both “mergers” can appear simultaneously, if the Goal is merely to experience a resource; in this case the Goal is the situation in which the aimed-for action of retrieving the Object has been performed. Either of the three may be targeted, and they are all conceptually different, but closely related.

It is possible for Stage 1-points to be “merged” in the same way. For instance, the need may simply be to understand a topic. In this case, the Subject and the Trigger are very closely related, but one remains a topic domain and the other a type of need. There may also be no further Source situation apart from a perceived need, whether it be topical or otherwise. In this case, Source and Trigger are intimately connected. Furthermore, all three may be “merged” in this way if the Source situation is simply one of a perceived need for information on a topic – a prototypical topical case.

It may look as if such pseudo-mergers shorten the paths by eliminating some of the potential targets. However, all potential targets are invariably present (in the sense that it is possible to distinguish them and discuss their differences), although they may be subject to pseudo-mergers that reduce the importance of the distinctions in such cases. The reason that targets tend to merge in this way will become clearer when we have explored Trigger and Aim foci in section 8.3.3. These foci seem to be at the heart of what user needs are.

It is tempting to see Objects as a step towards the Aim, and Aims as a step towards Goals, etc., so that everything becomes one timeline or process, but in spite of the account that has been given above of the intimate relationship between the search process and the various targets it is suggested here that it does not work like that. Rather, the possibilities of pseudo-merger govern the power relationship among targets and reflect the sense of an order between them (hinted at in the figures by the length of the arrows). The targets are not fundamentally sequential elements (hence the lack of arrows or connecting lines in fig. 8.9), but they do arrange themselves in this way which parallels a prototypical search process, due to the restrictions incurred by the possibilities of pseudo-merger.

8.2.4 Evidence of query types

As secondary evidence of the existence of these target types we may turn to the query length data again. When the query length distributions from L^{WID} are plotted in a diagram, an interesting distribution appears (fig. 8.10). There are clearly two different slopes involved, with S1 and S2 belonging to one and T1 and T2 to another. This leaves us to explain why S1’ belongs to the latter and S2’ to the former. One possible explanation is given in chapter 10 where the method employed for classifying the queries is given.

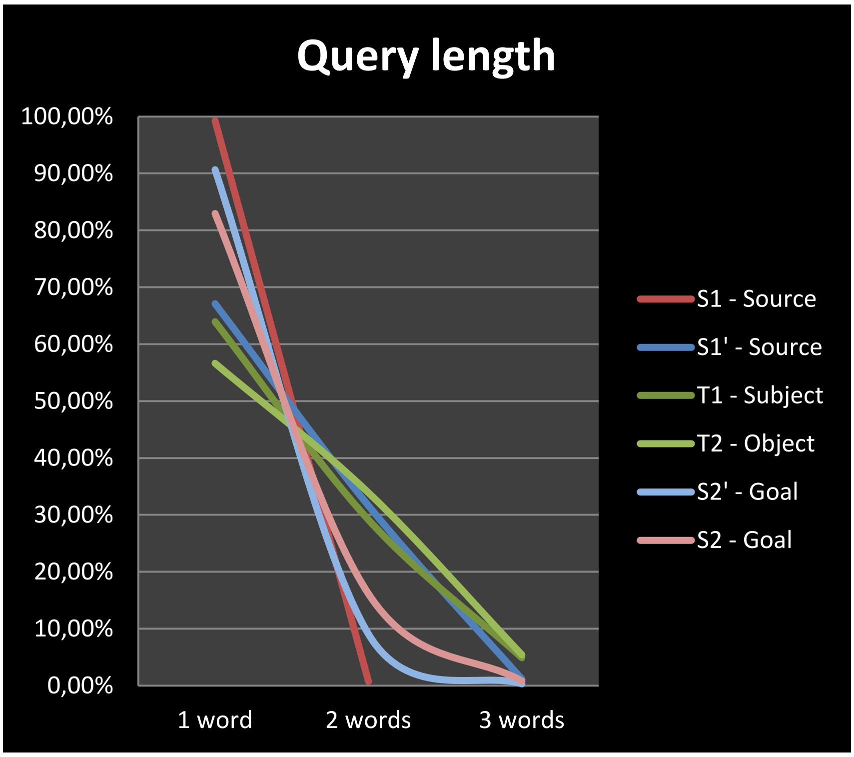


Fig. 8.10. Distribution of 1-, 2- and 3-word queries as percentage of the total number of searches referring exclusively to each of six main referential targets in L^{WID}.

If the data are plotted on a logarithmic scale, an inexplicable effect emerges, which deserves further study in future (fig. 8.11). Suddenly, there appear to be three different tendencies, one corresponding to the S1-S2 dimension, one for the T1-T2 dimension and one for the aspectualized S1'-S2' queries. The notable thing is the mysterious hump at 5-word queries occurring across all targets except S1' and S2'.

The dataset is far too small to conclude anything from this hump, but it is nevertheless highly suggestive that it should be present with all targets except S1' and S2'. Further research is needed on this point if an explanation is to be found. However, one suggestion deserves to be mentioned: It is possible that the hump represents a second slope starting at 5 words and contributing to a higher number of queries from that point on, and that this might represent the use of natural language expressions or some other type of query which is not normally possible with fewer than (approximately) five words because of the need to express certain syntactic relationships between items.

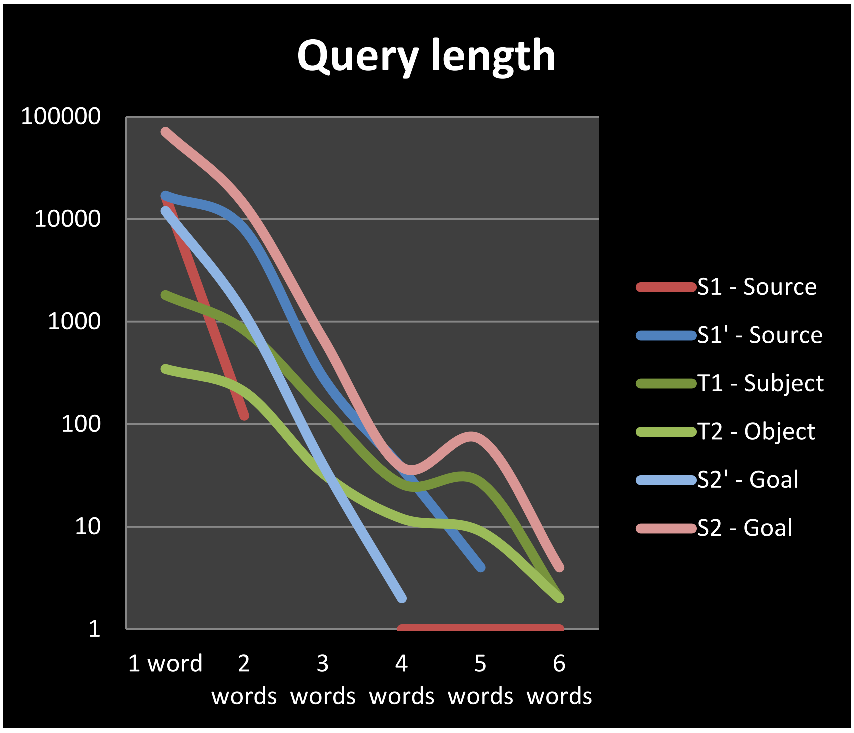


Fig. 8.11. Logarithmic plot of the absolute number of searches with lengths from 1 to 6 words. This time, there appear to be three tendencies corresponding exactly to the three dimensions involved. Note the mysterious hump at 5-word queries.

8.3 Recognizing a need

Typologizing over needs is difficult, partly due to the immense variability of need-generating situations, and partly because characterizing a situation in the first place is a never-ending task. Initially, we tried a simple solution based on responses, but in many ways such a typology says more about the nature of the system than the intention behind a given query. Attending to the user need that triggered the search is likely to be far more revealing. However, needs are many things. Objects, Aims, Goals and Objectives are subtypes of one way of looking at needs, while Subjects, Motivations, Sources and Triggers represent another. Summing up, the dimension from Motivation to Objective comes from within and is characterized by affective reflection and emotion. The dimension from Source to Goal, on the other hand, is fuelled by perception and awareness, and the one from Trigger to Aim by rational decision. Finally, the dimension from Subject to Object is characterized by planned action. There is some planning in the Trigger – Aim dimension as well, but it is mostly directed towards the problem and concerns ways of resolving it, whereas the planning that occurs in the Subject-Object dimension is very much characterized by a meta-awareness of the abilities and restrictions of the system and the kind of resources that are expected to be available, and also the search tactics that are likely to provide the best results.

The diversity of Source situations is one of the main reasons why the methods suggested in this thesis work best in narrow domains, where a distinction can fairly easily be made between the most common types of situations. Adopting the stage-based intent model and performing a *Target Analysis* is clearly much simpler (or in other words, is possible), as compared to attempting to classify logs according to actual differences in intention. Even so, the model is still very complex, and the above does not even treat the various facets that Subjects, Objects, Sources and Goals exhibit (cf. section 2.3). The result is a highly complex and sophisticated model with a multitude of beginnings and ends, and this reflects the diverse uses of search that can be observed. This section digs deeper into the workings of the model and exemplifies some of the processes covered.

8.3.1 Complex user behaviour

Users being people, they do not always follow the ideal model but show complex behaviour. Consequently, when we get to the analyses of the actual data we shall see that even with the full model in place there are some instances that seem not to be covered unless we recognize some specialized patterns of search behaviour. This should not discourage us, as long as this behaviour can be accounted for in terms of the basic model too. For instance, sometimes users plan sessions with several consecutive purposes. A user may be aware that if the first query is particularly successful or fortunately chosen the entire session may in fact be concluded with a single query, in that an intermediate Goal can be skipped and the ultimate one fulfilled. This is the case when there is an intermediate intention of checking whether a term is correctly spelled, for instance. If it is indeed correctly spelled, the user may then proceed directly to the task of gathering information on the topic in question, etc. This may be called a **compressed session**; two tasks are completed in one search, if and only if the first is successful. Another instance is navigation to a site, which is most frequently an intermediate Aim, but one that may likewise fulfil a more substantial purpose, which is also a natural outcome of the search process and hence an Aim. A user may aim to see a certain illustration for instance, but choose to aim for the page first. Furthermore, when people search for a product in order to find information about it with the ultimate purpose of buying such a product, this may also lead to a compressed session that ends in purchase even though the query in fact merely expresses a need for information, images, prices or specifications concerning products of a given type, if and only if the search for a suitable product is successful. Finally, so-called *teleporting queries* (Teevan, Alvarado, Ackerman, & Karger, 2004) are instances of compressed sessions in the same way as navigational ones.

The opposite phenomenon to compressed sessions, **optimistic Aims**, is very common with transactions as well as with other queries which are traditionally assigned to the transactional type, sometimes for exactly this reason. Optimistic Aims involve an Aim of completing a transaction or other action, such as cancelling a subscription, printing an invoice or executing a web-based program. In such cases, users often go for the ultimate Aim or Goal even though they are perfectly aware that it must necessarily require them to go through at least one additional step of clicking some button or link after having reached the proper site. Maybe they will even have to make a selection or fill in a form, but still they consider it as essentially one transaction for the purpose of searching. This might seem to justify the collection of all such queries under the umbrella category of transactional queries. However, optimistic Aims collapse Aim and Goal into one even though it is only possible to satisfy an intermediate Aim directly by search, and with our definition of Goals, it is not defensible to treat all Goals liable to optimistic Aims to be of the same type (such as *transactional*), as there is a clear difference between acquiring a product, collecting information via a web-based program, and executing an external action such as printing. The stage-based intent model treats these as separate Triggers, as the basic intentions are highly different. Consequently *optimistic Aims* are reduced from a type of Goal or target to a behavioural phenomenon.

8.3.2 Defining Triggers

Probably the dimension that is hardest to come to terms with is the one from Trigger to Aim. At the same time, it is one of the most important, being in fact what the Broderian classification seems primarily to attempt to get at, namely the difference between a need for an answer to a question, a fact, some information on a topic, or access to a particular resource. It is therefore worth looking a little closer at the various Trigger types. When thinking of Triggers it is important to be aware that they are not defined by any kind of linguistic differences (although they are typically accompanied by such differences), nor by the form of response that they anticipate. Triggers are of a kind separate from and independent of all of these considerations. While the Source is the local situation as perceived by the user, the Trigger represents the focal problem that the user decides to concentrate on solving with the view to ameliorate the problematic Source situation. The Subject is then a kind of anticipated solution – a stab at a topic which might provide the answer to the question, etc., if the system behaves in a helpful manner and in accordance with expectations. In other words, while Aims suggest the shape of an ideal outcome, Triggers define the shape of the “hole” or “gap” that needs to be filled or closed by reaching the Aim. Triggers seem to guide and define user behaviour to a large extent. The following is a summary of the commonest Trigger types as defined here.

- **Need-triggered search**: The user wants something which will satisfy a certain need, such as a craving for entertainment. (Motivation in focus)
    - **Serendipity-triggered search**: The user is not particular about the exact resource, as long as it is pleasing or relevant in some way.
- **Advice-triggered search**: The user wants advice on how to act in a given type of situation. (Source in focus)
- **Gap-triggered search**: The user wants the answer to a question, or is looking for a fact. (Trigger in focus)
- **Topic-triggered search**: The user wants to acquire knowledge on a certain topic. (Subject in focus)
- **Resource-triggered search**: The user wants to obtain something. (Object in focus)
    - **Acquisition-triggered search**: The user wants to acquire an object.
    - **Experience-triggered search**: The user wants to experience a site, a movie, a piece of music, etc.
- **Action-triggered search**: The user wants an action to be carried out. (Aim in focus)
- **Change-triggered search**: The user wants to enter a certain state. (Goal in focus)
- **Ability-triggered search**: The user wants to be able to achieve something. (Objective in focus)
- **Refinding-triggered search**: The user wants to refind a previously visited site or paragraph. (Any focus)
- **Appearance-triggered search**: The user wants to check the distribution, definition or appearance of a term or resource. (Results list in focus)

This list is not meant to necessarily be exhaustive, but the types shown seem to be among the most important Triggers encountered[179].

A perusal of the relevant literature soon reveals that it is common to distinguish between *fact finding* and *information gathering* queries (Sellen, Murphy, & Shaw, 2002; Kellar, 2006; Jansen, Booth, & Spink, 2008). This is reflected in the present account by the distinction between **topic-** and **gap-triggered** searches. Bhavnani *et al.* (2001) go further in distinguishing between *factual*, *sample* and *exhaustive* (or *in-depth*) search. It is true that there is a difference between merely needing a more or less random sample of relevant material and intending to gather all available information on a given topic, or even just a large amount. However, this difference may be more related to the size of the resource collection than the IR system, since one document may in some settings exhaust a topic within that collection. The present author considers *sample* and *exhaustive* search to be not targets but basic *search types*, though they may be explained by introducing facets as discussed in section 7.3.3.

The distinction between topic- and gap-triggered searches suggested here also replaces Rose and Levinson's (2004) *undirected* vs. *directed search goal*. For these authors, undirected goals are characterised by a desire to know anything or everything about a given topic, while directed goals represent cases where the user wants to learn something particular about the topic. These are two entirely different states with the potential to occasion different search behaviour. Rose and Levinson further distinguish between *closed ("I want to get an answer to a question that has a single, unambiguous answer"*) and *open ("I want to get an answer to an open-ended question, or one with unconstrained depth"*) *directed goals*. It is not entirely clear to me whether this really belongs with the Trigger. That question deserves looking into in future.

Note that a topic may be used in a query even when the search is not topic-triggered (as opposed to Subject-targeted). For instance, facts may be located and advice found by searching for information on a topic, if no better way is readily available. In fact, all Triggers allow topics to be used, and none require them. Thus, a topic-triggered search may be performed by navigating to a particular resource, or even by directing the search towards buying a book on the subject. It is only when the Trigger is *targeted* in the query that it appears on the surface in the expression of the need.

Advice-triggered searches are problem-oriented; people either describe their predicament or a possible solution. In practice, this may be achieved by locating information or getting answers to questions, but it is a distinct Trigger that occasions special search behaviour such as a strong tendency

[179] There are certainly many subtypes; these are mainly differentiated by the various facets of the model's endpoints (cf. section 2.3). For instance, topic-triggered queries may target the formal topic (understanding and using the concept correctly) or the semantic topic (understanding and acquiring knowledge on the topic), etc.

to describe the problematic situation rather than searching for actual solutions. Thus, while topic-triggered search has a strong focus on the Subject and gap-triggered search places great emphasis on the Trigger itself, advice-triggered search focuses the attention on the Source situation. However, it should once again be noted that this is independent of the choice of referential target.

The last Trigger type that focuses on Stage 1 is **need-triggered search**, which emphasizes Motivation. This is the kind of search in which the user does not care much about what resources are returned as long as they satisfy the need associated with the Motivation. **Serendipity-triggered search** is a sub-type of this. Serendipity usually means searching for no other reason than to be entertained; many sources call this *browsing search*. Note that in our model the drive to be entertained is separated from the actual behaviour of searching with no particular Goal, even though they frequently co-occur. Entertainment-motivated searches may of course be topic- or experience-triggered, and it is even conceivable that people may sometimes browse in this way for other reasons than entertainment. Thus, a potential target *may* be the focus of a Trigger, and *may* even be chosen as a target, but this is in no way certain. This provides for a highly flexible and complex model that is able to explain the great variety of observable behaviour.

Resource-triggered search focuses on the Object. The point is to collect resources or acquire goods. Among the subtypes, **experience-triggered** searches are closely related to a type noted by Rose and Levinson (2004) and called by them *resource searches*. This type is generally considered to belong with transactional queries. Indeed, one might think that locating an electronic video and being sent a physical one would be the same thing, but in fact the former is much closer to navigation than transaction, and it should therefore be considered a separate Trigger, if only a subtype. **Acquisition-triggered** searches are the prototypical transactional cases, but interaction with databases and sites does not always belong with this type of Trigger. Typically, users who perform acquisition-triggered searches describe the item that they want, rather than the situation of lack that triggered the search. For instance, going to a webshop that sells smartphones you do not usually describe your *smartphonelessness*, but rather the item that you would be interested in buying. Similarly, the distinction between **advice-triggered** and **change-triggered** searches focusing on the Source and the Goal, respectively, explains the difference in behaviour between users looking for the solution to a problem and those simply wishing to enter a new state that was noted in section 8.1.3; advice-triggered search causes users to focus more on the problematic Stage 1, while change-triggered search makes them focus on the ideal Goal situation.

Ability-triggered search is somewhat more abstract, and may possibly be less common than the other types. It has not been observed in the empirical data analysed here. What makes this type less straightforward is that while change-triggered search implies that the user wants to effect a certain change but does not basically care how this is achieved, having ability as a Trigger means that the user does not even care what the outcome is, as long as the long-term Motivation is gratified. This makes it somewhat similar to its opposite, the need-triggered search which focuses on Motivation. However, just as change-triggered search is advice-triggered search without the problematic situation, ability-triggered search is need-triggered search without the actual need. Thus, it implies browsing for potential benefits without the drive created by an existing concrete need. This sets it apart and undoubtedly makes it rare. It was added to the typology for the sake of completeness.

People regularly need to refind a resource that they have seen before (Elsweiler, Baillie, & Ruthven, 2011), and **refinding** is a special Trigger that tends to target actual resources rather than content or outcomes. Finally, some searches are simply performed in order to check the spelling of a term, the distribution or frequency of occurrence of a word, etc. These are examples of **Appearance-triggered search**.

8.3.3 Trigger and Aim foci

The choice of a target always implies a certain **resolution**, i.e. the Stage 2 corresponding to the target. If the target is itself located at Stage 2, the resolution is equal to the target, but if the target is Stage 1, the resolution differs. For instance, if the user chooses to target the Source situation, this implies that reaching a Goal situation where the problems characterizing that situation have been solved is the implied resolution. If the user targets the Subject, the resolution is the Object containing information on the topic. This is not to say that such a resolution will necessarily solve all aspects of the problem, or for that matter that the resolution will necessarily be realized; it is simply what the query implies. The importance of resolutions in this sense is that they tend to shape the user's conception of the problem, thereby strengthening the idea that led the user to choose a certain strategy in the first place. Thus, if a user chooses to target the Source, this implies a Goal resolution, and this will cause the user to envisage the solution in this way rather than starting to think of various relevant topics that one might also search for in such circumstances, and perhaps worse, the user who targets the Subject will tend to focus on finding the best topic or string possible, rather than thinking in terms of solutions to the initial problem as such. Indeed, resolutions are likely to be what triggers people to choose a certain target in the first place, and different people probably differ in their default choice of resolution; some will tend to think in terms of Goals or even Objectives, while others will be more focused on desired Objects and Aims. Paradoxically, while the former are perhaps more practical as regards solving the ultimate problem, the latter are more realistic when it comes to interacting effectively with the tools, at least at present. This begs the question whether tools are constructed in the wrong way, especially as the focus on topic search that is so prevalent (Ingwersen's *label effect*) is probably strongly augmented by experience with current tools and their deficits. The alternative answer is that computers are inherently incompatible with human problem-solving, and that consequently this schism is inevitable, as argued by Winograd and Flores (1986). The present author has no definitive answer to give in this dilemma.

As shown above in the list of common Triggers, each type has a focus in the form of a target that is at the centre of that Trigger. This **Trigger focus** is at the heart of user needs, because it defines (in certain less than definite ways) what the user needs most of all. When users perform advice-triggered searches, odds are that they will target the Source situation, and that the *resolution* specific to that dimension – the Goal situation – is what they care most about; they want to reach that situation with the least effort possible. If, on the other hand, someone performs a topic-triggered search, then he or she will most likely (but in no way necessarily) target the Subject and will need to find some information on that topic, i.e. the Object. The Aim will simply be to retrieve the relevant resources, and the Goal to have them. In this light, what really matters is the **Aim focus** – the possible target that is found in the opposite end of the dimension to which the Trigger focus belongs, and the possible targets "further up the line" are less vital, because they normally more or less echo the same need as viewed from various angles. This accounts for the pseudo-mergers in an elegant way; if the Aim focus is the Object, owing to a Subject-focused Trigger, then Object and Aim all but merge, and even the Goal is moulded by this purpose. On the other hand, when the Trigger focus is the Motivation, the Aim focus is the Objective, and the user is likely to target the Motivation and to be most centrally interested in having the need emanating from that Motivation satisfied. In this case, Object, Aim and Goal are entirely separate and may represent completely different steps to be taken.

It would of course be entirely possible to compile a list of Aim types to match the one given for Trigger types. However, as part of the general model, we shall prefer to consider the Trigger and Aim types to be mainly the result of the various Trigger and Aim foci and shall refer to those instead whenever applicable. This makes the framework more abstract and theoretical for better or worse, and so it is entirely understandable if people in general will prefer the more concrete Trigger types when discussing query classification, which is why they were listed in full above.

8.3.4 The essence of a need

The Trigger (and Aim) foci make up a hierarchy similar to the HRC discussed in Part II, which we might call the **Hierarchy of Referential Foci** (**HRF**). Similarly, the targets themselves may be thought of as constituting a hierarchy that parallels the HRD hierarchy of relevance dimensions. This we may call the

Hierarchy of Referential Targets (HRT). The HRF is at the heart of the experience of a need, and just the relationship between the HRC and the HRD could be considered the essence of relevance (cf. section 6.3.2), so the relationship between the HRF and the HRT is arguably the very essence of needs. It was noted in connection with relevance that the order of the elements in the two hierarchies differs due to the human propensity to notice the most immediately relevant first (focal relevance).

As with the HRC, the HRF is likely to have this same skewing, so that a Subject focus is the way of considering a problem that springs most easily to mind when approaching a LAT. This would explain the label effect as an entirely natural part of human perspective in parallel with the priviledged status of topicality in relevance assessment (fig. 8.12).

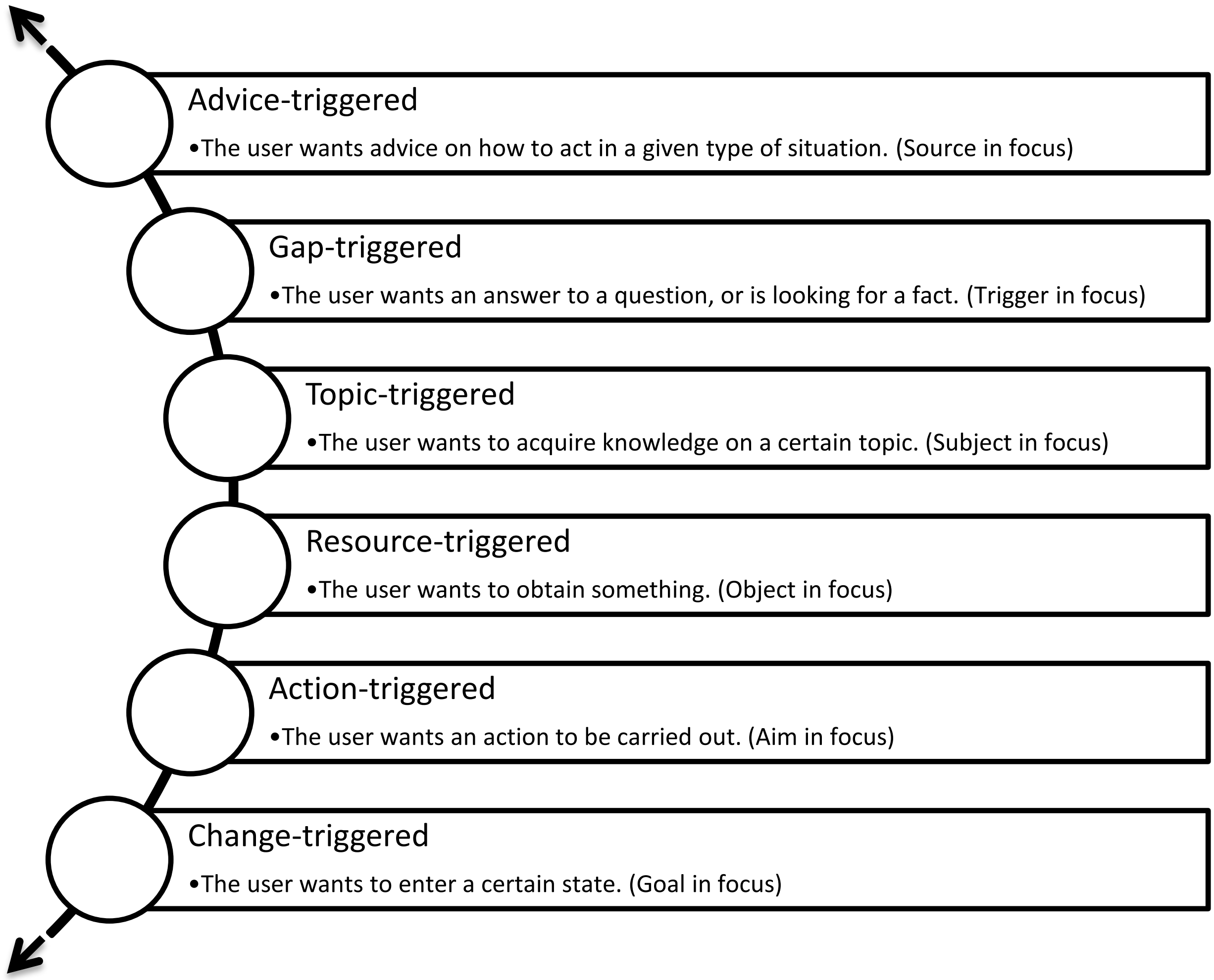


Fig. 8.12. The hierarchy of referential foci (HRF). The order is inferred from the HRC based on the Schützian spheres of context and confirmed in part by the existence of the label effect. This hierarchy would ultimately have to be confirmed by psychological experiments. Motivation- and Objective-triggered types were excluded for reasons of space. They rarely occur because that would imply that all other possible solutions had been exhausted since they occupy the absolute periphery.

The partly reversed order of the HRF implies that if the user is after topical information the other possible foci need not even be considered; the user can proceed directly to performing a topical search without bothering about motivations or even problematic situations. If, on the other hand, the user has a problem then topical solutions are still considered first, but may be overridden in favour of a Source focus. The extact order of foci would have to be confirmed by psychological experiments. It is probable that the two halves of the hierarchy proceed in parallel from a common center where topics and resources are considered to the peripheral Motivations and Objectives. This incidentally also explains the rarity of Motivation and Objective as target; according to the HRF they are only ever considered if all other possibilities have been exhausted.

It may strike the reader that there are eight foci and only six relevance dimensions, and indeed it is possible – even probable – that two additional relevance dimensions may be recognized in future. If so, these would be a highly abstract dimension corresponding to the Trigger between topicality and utility, which would deal with the question whether the resource provided the desired kind of input, and a similarly abstract one corresponding to the Aim between informativity and practicality whose job it would be to determine whether the resource evaluated was of an appropriate kind (such as text vs. image or video resources). Most likely, this would mean adding two additional facets as well – an ontological facet representing in some sense the type of resource, and a telic facet representing the type of content or effect offered. Far from disproving the suggested framework, the need for additional dimensions would bring the opportunity of seeing how easily the model can be adjusted in this way, and how much of an advantage this is to having to construct a new model from scratch. The consequences of inserting these particular two dimensions have not been thoroughly worked out at this point; there is still work to be done.

9 Studying query variation

> *Since search queries are a fundamental part of the information retrieval task, it is essential that we interpret them correctly. However, the variable forms queries take complicate interpretation significantly. We hypothesize that elucidating the grammatical structure of search queries would be highly beneficial for the associated information retrieval task.*
>
> (Barr, Jones, & Regelson, 2008)

The fact that the average query is proverbially short and ambiguous has occasioned a substantial literature on how to disambiguate terms[180]. The present thesis does not aim to perform this kind of *semantic* query disambiguation. Rather, it seeks to better understand the user's intention in the first place, to the extent that this is possible by attending to the linguistic details of queries and to contextual clues. Queries show all kinds of variation in their expression; changing word orders, various inflected forms of words, and different word choices by different users and in different situations, and at present very little is known about why such variation should occur and what intentional differences it reflects, if any. Precious little has been written on the subject of the linguistic structure of queries; since the seminal paper on query syntax by Barr *et al.* (2008), work has been kept up notably by Xiao Li and associates (Manshadi & Li, 2009; Li X. , 2010), but being a rich source for further research, these papers represent very different takes on the problem, some being more semantic than syntactic in their approaches.

We have seen in the course of the previous chapters that queries reflect many different intentions, even apart from the actual topics, needs or triggering situations involved. The question is how this is coded in actual expressions.

What are the relationships between form and intention in a query?

[180] See Bhogal *et al.* (2007), among countless others.

The reader will hardly be surprised to find that there is no simple relationship. Rather, there is a multitude of tendencies and principles which are frequently at conflict. The main types of coding used in queries are morphology and syntax, including word order and phrasal structures. Modification is rife, and the object relation is in frequent use, both with verbs and in nominalized constructions. What is mostly lacking is the subject relation, along with other verb-noun relations normally expressed through case and adpositions. The remainder of this chapter explores some of the linguistic characteristics of queries, based on data from the two log files L^{TDC} and L^{WID} introduced in Part I. No attempt is made here at providing an exhaustive account of query variation or giving any in-depth treatment of the difficulties associated with it; we shall mostly be using the data as an illustration of the enormous amount of variation that occurs in a search log and that any framework for IR research must be able to deal with. We shall, however, be keeping an eye open for phenomena which might be correlated with the choice of target, in order to substantiate the suggested framework.

9.1 Query morphology and syntax

Morphology is the study of the different forms a word may assume, such as singular or plural for nouns, and present, infinitive or imperative for verbs. In a primitive IR system such as barely exists today, such differences present a problem because only results that contain the exact form searched for are returned. For this reason, most systems currently employ the process known as **stemming**, and many achieve this by simply **conflating** all forms of a given lexeme, i.e. by removing all endings under the surface so that only the stems remain[181]. This eliminates the problem of not finding documents containing other forms than the one searched for but not an exact match. However, it might actually be the case that different forms convey different meanings, and hence that the user would not necessarily be interested in documents containing other forms, or conversely, it is even conceivable that users who provide a certain form *primarily* need documents containing a certain other form of the same word. For instance, it is a perfectly reasonable assumption that a user issuing a query such as *dog* using the basic form would be more likely to be interested in documents containing occurrences of the generic term *dogs* than those involving the singular *dog*. If the endings are kept in system memory, different weights can be assigned to the various forms used in order to take care of such cases. Alternatively, the system may be based on a complete dictionary that recognizes all forms of words without having to resort to crude stemming[182]; this technique is known as **lemmatization**[183]. This approach is highly successful in

[181] An alternative is inflectional and derivational query expansion where all forms of a lexeme and its derivatives are automatically added to the query string containing one.
[182] Detailed experiments have shown that automatic stemming does not improve performance in web search (Harman, 1991; Kraaij & Pohlmann, 1996). The number of queries that benefit from stemming is in fact similar to

enterprise search, but problematic in web search because in principle the dictionary must contain all occurring words and this is not feasible with such a large and diverse corpus.

Query syntax is no less prevalent than morphologically marked forms. The term is most commonly associated with formal languages designed to help users retrieve exactly what they need from a database, and it may come as a surprise to many that queries often include linguistic syntax, seeing as they usually comprise no more than two or three words (cf. section 5.2). Yet, a large proportion of search queries in the logs analysed do contain syntax. We even find two kinds of syntax; on the one hand, syntactic constructions from natural language such as noun and verb phrases are used freely[184], and on the other you find special constructions specific to queries developed through exposure to the workings of search engines.

> [Q]ueries exhibit their own partially predictable and unique linguistic structure different from that of the natural language of indexed documents (Barr, Jones, & Regelson, 2008, p. 1022)

Whenever the term *query syntax* is used in the course of this thesis, it is meant to be understood in the linguistic sense, as a complement to *query morphology*. Note also that the term *phrase* will be used with its linguistic meaning of a head (typically a noun or a verb) with its modifiers. In the IR literature, it is regularly used in two other senses of *fixed phrases* (or multi-word terms) and *statistical phrases*, i.e. words which co-occur in the dataset, typically with no regard to syntactic (or semantic) structure (Arampatzis, van der Weide, van Bommel, & Koster, 2000).

The presence of syntax is to a wide extent due to the necessity for information structure. Syntax and information structure are thus closely related issues. One syntactic rule might for instance be thought to dictate that the most important element be mentioned first, and it is therefore natural to ascribe

the number of queries which deteriorate. Popovic and Willet (1992) found that stemming did indeed improve precision significantly when employed for a Slovene corpus, but when repeated on an English document set the benefits evaporated. There may thus be some effect of language type, and especially of morphological complexity (Ingwersen & Järvelin, 2005a, p. 154). The benefit that comes with enterprise search is that automatic stemming is not necessary; lemmatization dictionaries are less costly both in performance and maintenance, and relations and weights can be updated manually. In this way, many of the disadvantages associated with stemming in a web search environment where it would be too costly to introduce many exceptions and downgraded relations between inflections and derivations are minimized.

[183] Note that the terminology is not entirely standardized: Some treat *lemmatization* and *conflation* as different types of *stemming*, others consider *lemmatization* and *stemming* to be *conflation* techniques, yet others more or less equate *stemming* with *conflation* and consider *lemmatization* to be an alternative approach. This may easily create misunderstandings, and consequently claims as to the effectiveness of stemming should be qualified by specifying what that notion is meant to cover.

[184] In particular, noun, adjective, verb and prepositional phrases.

more weight to the first (lexical) item in a query than to others. Some systems also attempt to identify phrases (especially noun phrases) in order to improve the semantic match between queries and documents (Bergsma & Wang, 2007). However, the primary concern in such studies is in fact to identify multi-word terms; apart from this, syntax is generally ignored. The reason is of course that natural language processing is extremely difficult, and often the rewards do not compare with the efforts and investments required. Consequently, what is suggested here is not a complete linguistic reconstruction of a query's underlying structure but a simple parsing which allows the system to recognize constructions that reveal facts about the intention. If morphology and syntax is found to vary with the choice of target, then this information can be exploited in automatic target recognition. Apart from this, syntactic parsing allows recognition of semantically equivalent phrases, which is of great advantage to the users of most systems, given that, as we shall see shortly, they in fact use a surprising amount of complex phrases and constructions which do not necessarily match their counterparts in documents exactly[185].

	Singular	**Plural**
Nominative, indefinite	*telefon* 3,129 (71%)	*telefoner* 990 (22%)
Genitive, indefinite	*telefons* 6 (0%)	*telefoners* 0 (0%)
Nominative, definite	*telefonen* 292 (7%)	*telefonerne* 3 (0%)
Genitive, definite	*telefonens* 13 (0%)	*telefonernes* 0 (0%)

Fig. 9.1. Morphological variation in the lexeme *telefon* 'telephone' as it occurs in the TDC search log. The number given for the basic form excludes all occurrences as first part of compound words such as *telefonsvarer* 'voicemail'. The numbers represent distinct occurrences during one year 2010-2011.

9.1.1 A question of form

It might be thought that not many non-basic forms are used in queries anyway, and that stemming is primarily of use in the indexation of documents rather than in that of queries. However, while it is certainly true that, all other things being equal, there is a preference for shorter and more basic forms of lexemes when formulating queries, all other things are in fact not equal, and the number of non-basic forms turns out to be surprisingly great. Search logs often reveal morphological variation, such as

[185] Arampatzis *et al.* (2000) suggest one way of dealing with expressive variation, while Ingwersen and Järvelin (2005a) discuss the issues involved more generally.

the same noun phrases in singular and plural form, or different forms of the same verb. Fig. 9.1-9.2 exemplify this by showing the variation in form in the lexemes *telefon*, 'telephone' and *afbestille* 'cancel (a subscription)' respectively, in L^{TDC}.

Form	*Example*	*Frequency*
Imperative	*afbestil*	43 (25%)
(Active) infinitive	*afbestille*	33 (20%)
Active present	*afbestiller*	0 (0%)
Passive present or infinitive	*afbestilles*	2 (1%)
Past participle	*afbestilt*	0 (0%)
Nominalization	*afbestilling*	91 (54%)

Fig. 9.2. Morphological variation in the lexeme *afbestille* 'cancel' / *afbestilling* 'cancellation' as it occurs in L^{TDC}. For the purposes of searching they may be considered to be variants of one lemma.

Most search engines still ignore the morphological information and effectively treat morphological variants as synonymous by conflation. Some people may even be surprised that this kind of variation occurs at all. One thing is apparent from the data analysed in connection with this project; people do regularly utilize such forms as imperatives, passives and definite nouns in search, and the pressing question is *why*. There are various reasons for such uses, some of which will hopefully be clearer after reading this thesis. The proportion of imperatives to infinitives and verbal nouns, for instance, is the subject of a study performed in section 9.4, where a possible explanation for the variation is also suggested.

The idea of conflation of forms is based on the assumption that all morphological variants are more or less equivalent, and while forms of the same root do indeed represent *related* intentions, the basic assumption of equivalence is fundamentally mistaken; a text on *dogs* may in fact be vastly more useful to someone interested in dogs than one containing the phrase *the dog*. Similarly, if the user enters the singular form (and especially if the query *a dog* should be chosen) it is most likely that he or she does not intend to find general information on dogs at all. Rather, possible intentions might be buying one, or locating information on how to deal with a certain situation involving an instance of the type referred to, as in the following imaginary example.

9.1	**sick dog** {S1: sick dog$_{n.sg}$}	*

It is thus no coincidence that requests for advice on how to deal with problematic situations tend to use the singular rather than a generic plural, as exemplified by the small-scale study shown in fig. 9.3. It is remarkable that not a single one of these queries uses a plural topic – they all refer to one instance, either as a non-specific referent in a certain type of situation or as a specific occurrence in the situation in question. Clearly, if morphology is related to intentions in this way, LATs need to pay attention to what is actually said, in queries as well as in target documents, rather than simply discarding the morphological variation through conflation.

Sometimes, seemingly innocent changes in the expression of a query may represent entirely different intents. There is likely, for instance, to be a difference in intention between the following two queries.

9.2	**nye telefoner** {T1: nye telefoner$_{n.pl}$}	'new telephones'	TDC (8)
9.3	**ny telefon** {S2: ny telefon$_{n.sg}$} / {S1: ny telefon$_{n.sg}$}	'new telephone'	TDC (84)

There is a morphological difference between the two; one is in the plural and the other in the singular. The former is likely to reflect a wish to collect information on or images of relatively new telephone models, while the latter seems to represent a user's desire to find and buy a new telephone, or perhaps to find information on how to buy one, etc. The much higher frequency of the singular form of course has other, less interesting, reasons apart from the difference in meaning, such as brevity of expression and basicality of the uninflected form, but the point to notice here is that the plural form occurs at all. If there was no meaning difference between singular and plural query terms, such forms would not be expected to appear – Why would anyone choose to use a plural form? The choice must hence represent a difference in intention.

Query	Example	Frequency
$BROKEN_{adj}$	**i stykker** 'broken'	2
$BROKEN_{vb}$	**gået i stykker** 'broken'	1
DEVICE $BROKEN_{adj}$	**telefon i stykker** 'telephone broken'	26
DEVICE $BROKEN_{vb}$	**telefon gået i stykker** 'telephone broken'	2
DEVICE HAS_BROKEN	**telefon er gået i stykker** 'telephone has broken'	2
DEVICE-DEF $BROKEN_{vb}$	**telefonen gået i stykker** 'the telephone broken'	1
DEVICE-DEF IS_BROKEN	**telefonen er i stykker** 'the telephone is broken'	2
POSS DEVICE IS_BROKEN	**min telefon er i stykker** 'my telephone is broken'	3
POSS DEVICE HAS_BROKEN	**min telefon er gået i stykker** 'my telephone has broken'	3
-	**kabel gravet i stykker** 'cable dug through'	1

Fig. 9.3. Analysis of problem-reporting queries containing the phrase *i stykker* 'broken' in L^{TDC}. The queries were normalized by grouping them according to their structure. Category memberships are as follows: DEVICE = {mobil, fastnettelefon, mobil, rrouter, iphon, telefon, email, iphone, simkort, sim kort, digital modtager, skærm, e-mail, display, fjernbetjening, home trio box, display, tkf, tdc boks, modem, taletid simkort, fjernbetjening til hometrio}; $BROKEN_{adj}$ = {i stykker, istykker}; $BROKEN_{vb}$ = {gået i stykker}; IS_BROKEN = {er i stykker}; HAS_BROKEN = {er gået i stykker}; DEF = (definite suffix *-en*); POSS = {min}. Full dataset in Appendix E1.

Even the meaning of the word *new* changes significantly from '(models) new these days' in the plural to '(phone) new to me' in the singular, and these are potentially entirely independent timescales[186].

[186] With animates there is a lexical distinction in English between *young* and *new* to distinguish these two meanings, but with inanimate things English speakers make do with the one word *new* (Taylor J. R., 2003, p. 28). The fact that there is only one word available to English speakers for inanimate objects does not mean that the difference does not exist, however.

Another example is *new passport* vs. *new passports*. The former would generally be more likely to be used in looking for information on how to renew your passport, while the latter is presumably more likely to occur in a situation in which the user has heard that a new passport design with augmented security has been implemented and information on the consequences of this is sought. There is no way that we can tell with certainty which meaning was intended in a given query, but if it can be assumed that occurrences in the singular are more likely to refer to something that is new to the user, whereas the plural signals that the telephone model is new, then this could be exploited by search systems. This would help model the intention of the user as accurately as possible.

In the interest of being able to discuss further possible relationships between form and intention, we shall now very briefly sum up some of the linguistic phenomena discussed in Alfort (2009). A noun in its most basic form simply represents a concept without any reference to actually occurring individuals, or even potential individuals which might conceivably exist in some mentally constructed situation. Such individuals require separate construction in a linguistically portrayed situation. When this occurs, the expression is elevated from the so-called *structural plane* where empty template situations are constructed from concepts to the *actual plane* where situations are portrayed by combining templates and filling in slots (Langacker, 2000). On the structural plane, there are no number distinctions, because referents are not *realized* in a portrayed situation. For countable entities, the plural is the unmarked number used for such unrealized terms. That is why generics are in the plural if they refer to countable entities. Plurals on the structural plane are *unbounded*, i.e. they do not refer to any quantity but merely to an unspecified mass of countable entities of a certain type. On the actual plane they are *bounded*, because they refer to actual or imagined countable entities. Since the basic form is underspecified it can only be used in actual reference by applying a generic or non-specific perspective, which is exactly what happens in topical queries.

Whenever a query term represents something that the user wishes to find something of, the referential form used is the non-specific, and when it refers to something of which the user wants some examples a specific form is used. In a topical query, on the other hand, the generic form is chosen. Thus, generics, non-specific and specific forms correspond to different query functions. In general, one might hypothesize that generic plurals would tend to represent topics, whereas indefinite but specific singulars would most frequently represent entities occurring in problematic situations (*what to do in case of a…*), and non-specific singulars entities that the user wants (*I want to buy a…*, *I want to see pictures of a…* etc.). This would make sense in terms of extensional reference, and would furthermore correspond to the use in natural language constructions. Whether users of search engines do actually honour such principles remains to be studied in detail, but there would be predicted to be a *tendency* towards adherence, if only as one among many others. Whatever the outcome of such future in-depth studies, it is certainly clear from the small-scale studies performed during this project that the distribution of singular and plural use is far from random, and that extensional reference does indeed play an important role. Some common examples of variation between singular and plural in L^{WID} may be seen in fig. 9.4.

Term	Searches using plural	Searches using singular	% plural
christmas tree	20	0	100%
medical device	26	0	100%
sow	18	0	100%
hot job	17	6	74%
child	38	18	68%
operation	25	14	64%
hotel job	33	19	63%
wage	20	15	57%
polymer	58	45	56%
composite	33	27	55%
flower	16	14	53%
material	31	31	50%
pig	232	267	46%
strawberry	40	64	38%
job	47	87	35%
event	36	69	34%
pharmaceutical	35	82	30%
animal	26	63	29%
cow	56	156	26%
horse	16	48	25%
fruit	15	60	20%
student job	58	242	19%
textile	17	85	17%
hotel	111	656	14%
restaurant	38	258	13%
salary	19	122	13%
farm	109	1298	8%
internship	23	619	4%

Fig. 9.4. Examples of singular vs. plural distribution in L^{WID}.

The words in this list belong to several semantic fields related to the jobs sought. The words *hotel*, *restaurant* and *farm* occur close together at the bottom of the list because in any one job you normally only interact with one of these buildings (unless you intend to build hotels etc., cf. chapter 10). It is therefore natural for people to refer to them in the singular. On the other hand, *medical device*, *child*, *polymer*, *composite*, *material* and *pharmaceutical* represent items that are commonly interacted with in pluralities. Belonging to a wider and more varied semantic field they are distributed over a much wider section of the scale, but mostly above 50%. *Pharmaceutical* is lower than the others, presumably because it is also an adjective, which does not have a separate plural form, and probably also because of its length, as words which are already long appear to be less likely to be weighed down further by the addition of suffixes. *Material* can also be an adjective, and possibly a mass noun, which would tend to lower the proportion of plural forms. Animals too are typically interacted with in pluralities, but even so they are thought of as individuals in a way that objects are not, especially as regards the larger animals. Consequently, *horse*, *cow*, *animal* and *pig* cluster below 50%, while *sow* is inexplicably represented in plural form only. It may seem surprising that children should not be individualized in this way, but in fact people who work with children tend to care for groups of children rather than individuals, whereas horses are often cared for individually.

Another tendency which may well be at play here is the reduction of compounds; while *pig* could easily be used with the intention of finding documents on pig farms and the like, *child* does not occur in similar compounds because there are special words for such institutions (cf. **child school*, **child teacher*). The last semantic group that we shall focus on is the one constituted by *christmas tree*, *flower*, *strawberry*, and *fruit*. These do indeed cover most of the scale, but the distribution is still not completely random. Firstly, *fruit* can be considered a mass noun, and therefore it may frequently occur in its singular form. Secondly, *strawberry* is longer than *fruit* and has an irregular plural ending – both facts that count against its being used in plural form.

Morphology may thus reflect user intentions in several ways, either because users anticipate the forms which are likely to occur in the resources that they want to find or because certain forms have stabilized into conventionalized ways of formulating certain kinds of requests. In some cases, the forms chosen may even reflect intentions more directly. This could be the case with some imperatives vs. other verb forms, as well as the distinction between the generic plural and other nominal forms; imperatives code requests and as such are likely to represent the actual aimed-for action rather than a topic label, for instance (which would normally be in nominalized or infinitive form), and while generics are typical in topic search, definite forms in all probability represent a wish to refind the string in a text, and indefinite singulars might conceivably be prevalent in descriptions of problematic situations where the user needs advice[187]. In other words, if these hypotheses hold water, we would expect *new cars* to be a topical request for information on new car models (Subject-targeting search), *new car* to have been issued by someone looking for a new car (Goal-targeting) or possibly wanting information on how to deal with one (Source-targeting) and *the new cars* to be a request for any documents containing that exact string (Object-targeting). Of course, it should not be expected that such principles would be obeyed in every case (for instance, the two indefinite noun phrases might just as well have been intended as complete strings to search for, and the singular form could just as well have been "meant" as a plural but not realized as such due to a predilection for short or basic forms), and so the hypotheses would be proven even by the slightest significant tendency for them to be followed more often than not, provided that the exceptions can be accounted for by regular principles in the form of countertendencies. We shall look at some of the tendencies involved in query formulation in section 9.4, and also study some of the common countertendencies that interfere with regularity, and we shall find that the observations are indeed compatible with, and even suggest, such correlations between form and intention.

9.1.2 Phrasal structure

A prototypical topical query consists of a single noun phrase and nothing else. Indeed, an automatic analysis of L^{TDC} revealed that at least 78% of the queries issued are of this form[188], while a manual analysis of the 1000 most frequent queries showed that among these 86% were simple noun phrases. In web search too, Barr *et al.* (2008) found that proper nouns constituted 40% of the query terms in their dataset, and a further 31% were common nouns, which is a roughly similar result. They also noted that the majority of queries are noun phrases rather than unstructured collections of terms. In their sample of 222 hand-labelled queries 70% were noun phrases. However, while they count the query *free mp3s* as a noun phrase, they do not recognize *mp3s free* as being a syntactic unit; rather they assign this type to a category called by them *word salad*, which accounts for a further 8%. We shall see later that this type appears to utilize grammar too; it simply follows syntactic rules which are absent from natural language English (NLE). Still, such structures probably do not count as single noun phrases but are undoubtedly complex (to the extent that they are not mere scrambled versions of simple noun

[187] These preferences can be overridden in cases of what we have called *compressed sessions*. For example, a user may enter a generic term as a topic in order to find a type of resource that he or she wishes to experience. Here, two Aims are concatenated, and if the first is successful the second becomes the ultimate Aim.

[188] The actual percentage is likely to be even higher due to nouns not recognized as such by the system (especially proper nouns referring to products and models, etc.). Even so, they represent only about half of the distinct queries, as shorter queries are naturally less diverse than longer ones. All automatic part-of-speech tagging used in this thesis was achieved using the powerful Ankiro Phonetic Spellcheck and Ankiro Thesaurus in combination, with Danish as the primary language for L^{TDC} and English as a secondary language. The opposite configuration was used for L^{WID}.

phrases, which of course is likely to be the case in some instances, though we suggest that this is not the main source of such constructions as they appear to have specialized meanings).

Noun phrases are not the only phrases occurring; a variety of structures are found.

> [T]he linguistically meaningful phrases which may be considered as retrieval terms are at least the noun phrase including its modifiers, and the verb phrase including its subject, object and other complements (Arampatzis, van der Weide, van Bommel, & Koster, 2000)

Barr *et al.* find 1.4% verb phrases and 0.5% full questions in their dataset, and de Lima and Pedersen (1999, p. 145) note that web queries are "often noun phrases, occasionally sequences of keywords, and even more rarely full sentential expressions". This is true of the enterprise search logs studied here as well, except that verb phrases are also of surprisingly frequent occurrence. Such phrases typically consist of an infinitive or imperative followed by an object noun phrase or a postposed prepositional phrase representing an oblique argument, to put it in linguistic jargon. Nominalized verbs are likewise common with a following prepositional phrase. Furthermore, as also noted by de Lima and Pedersen, the noun phrases used include some rather complex ones, sometimes involving several prepositional phrases in succession. A few examples from L^{TDC} follow.

9.4	**problemer med mail til hotmail** 'problems with e-mail for hotmail' {S1: problemer$_{n.pl}$ med$_{prep}$ mail$_{n.sg}$ til$_{prep}$ hotmail$_{n.prop}$}	TDC (5)
9.5	**overførsel af filer til homedisk** 'transfer of files to homedisk' {T1: overførsel$_{v.nom}$ af$_{prep}$ filer$_{n.pl}$ til$_{prep}$ homedisk$_{n.prop}$}	TDC (4)
9.6	**ændre abonnement fra mixmedia til click** 'change subscription from mixmedia to click' {T1: ændre$_{v.inf}$ abonnement$_{n.sg}$ fra$_{prep}$ mixmedia$_{n.prop}$ til$_{prep}$ click$_{n.prop}$}	TDC (4)

9.7	**ringe til udlandet med mobil** ‘call (to) abroad with mobile (phone)' {T1: ringe$_{v.inf}$ til$_{prep}$ udlandet$_{n.sg.def}$ med$_{prep}$ mobil$_{n.sg}$}	TDC (3)
9.8	**ringe med mobil til udlandet** ‘call with mobile (phone) (to) abroad' {T1: ringe$_{v.inf}$ med$_{prep}$ mobil$_{n.sg}$ til$_{prep}$ udlandet$_{n.sg.def}$ }	TDC (2)
9.9	**spærring af opkald til mobil fra fastnummer** ‘blocking of calls to mobile (phone) from landline number’ {T1: spærring$_{v.nom}$ af$_{prep}$ opkald$_{n.pl}$ til$_{prep}$ mobil$_{n.sg}$ fra$_{prep}$ fastnummer$_{n.sg}$}	TDC (2)
9.10	**overførsel af data fra mobil til computer** ‘transfer of data from mobile (phone) to computer’ {T1: overførsel$_{v.nom}$ af$_{prep}$ data$_{n.sg}$ fra$_{prep}$ mobil$_{n.sg}$ til$_{prep}$ computer$_{n.sg}$}	TDC (2)
9.11	**Priser på opkal til udlandet med taletidskort** ‘Prices on [calls] (to) abroad with prepaid card’[189] {N2: Priser$_{n.pl}$ på$_{prep}$ opkal$_{n.sg/pl}$ til$_{prep}$ udlandet$_{n.sg.def}$ med$_{prep}$ taletidskort$_{n.sg/pl}$}	TDC (1)

These phrasal structures are especially interesting when different constructions with the same semantics occur side by side, as they often do in the logs.

9.12	**oprettelse af fastnet** ‘establishment of landline’ {T1: oprettelse$_{v.nom}$ af$_{prep}$ fastnet$_{n.sg}$}	TDC (53)
9.13	**fastnetoprettelse** ‘landline establishment’ {T1: fastnetoprettelse$_{v.nom}$}	TDC (3)

[189] Square brackets in the translation indicates that the word in question is not correctly spelled but has been interpreted in the manner shown.

The prepositional phrase in ex. 9.12 can be reduced to a nominal compound with no change in meaning. Thus, unlike the other types of variation, such differences do not seem to differentiate intentions; the challenge here is to identify synonymous constructions and conflate them while keeping constructions that represent different intentions or variations on an intention apart.

9.1.3 Intentional word order

Perhaps the most obvious cases of syntactic variation in queries are word order differences as in the following examples from L^{TDC}.

9.14	**iphone mail** {T1: (iphone $mail_{n.sg}$)} / {T1: ($iphone_{n.prop}$) ($mail_{n.sg}$)}	TDC (361)
9.15	**mail iphone** {T1: ($mail_{n.sg}$) ($iphone_{n.prop}$)}	TDC (263)

We should ask ourselves whether such variation reflects a difference in intention or whether the order is entirely fortuitous. The reader will not be surprised to find that it will be argued here that the difference does indeed reflect intentions, although a certain degree of arbitrariness should always be expected. Ex. 9.14 is closely related to compound words; in fact, this particular example may alternatively be interpreted as such (in spite of the fact that compounds must ordinarily be written solid in Danish). Certainly, with an adjective as the first term the construction is entirely natural even outside queries.

9.16	**bærbar mac** 'portable mac' {T1: (bærbar $mac_{n.sg}$)} / {T1: ($bærbar_{n.sg}$) ($mac_{n.sg}$)}	TDC (1)
9.17	**mac[190] bærbare** 'mac portable/laptops' {T1: ($mac_{n.prop}$) ($bærbare_{n.pl}$)}	TDC (1)

[190] It is not entirely clear whether this is a proper noun in this instance or a common noun, and if the latter, whether it is a singular or an abbreviated plural. All of these interpretations are possible.

9.18	**hvid iphone** 'white iphone' {hvid $iphone_{n.sg}$}	TDC (48)
9.19	**iphone hvid** 'iphone white' {[$iphone_{n.sg}$] [$hvid_{n.sg}$]}	TDC (40)

The queries in ex. 9.15, 9.17 and 9.19, on the other hand, use special query syntax in that a postposed modifying expression is added to the noun phrase in focus. Ex. 9.16 is most likely to be interpreted as consisting of a single noun phrase, although it is possible to consider *bærbar* a noun (meaning 'laptop') and *mac* a proper noun, resulting in a different syntactic construction. The same is certainly true of ex. 9.18. Ex. 9.17, on the other hand, consists of a noun followed by what is either a plural adjective or a plural noun, and ex. 9.19 of a noun followed by an adjective. This cannot be one noun phrase, and the construction does not occur in NLE.

In natural language, the meaning behind ex. 9.15 would be likely to be expressed as *MAIL FOR IPHONE* or something similar, though perhaps the user's intention would be more properly covered by an expression such as *HOW DO I READ MY MAIL ON MY IPHONE?* On the other hand, ex. 9.14 would be more likely to be paraphrased as *TELL ME ABOUT MAIL ON IPHONE*, or even *TELL ME ABOUT IPHONE MAIL*. Thus, what initially may have appeared to be two synonymous expressions actually turn out to demand slightly different paraphrases, even in the very simplest of cases. The difference may be subtle, and many situations may not require a choice to be made between them, but still there is a difference and this may potentially be exploited for distinguishing between needs in at least some situations.

It is in no way unexpected that syntactic differences such as choice of word order should reflect intentions; for instance, it is only natural for a request for something to involve an initial indication of the kind of object desired, followed by a more precise specification as in the following utterance.

9.20	I would like to look at a Toyota. It should have five doors.

When transferred to a search situation, this is typically reduced to a simpler expression such as the following.

9.21	**Toyota five doors** {T1/S2: [Toyota$_{n.sg}$] > [five doors$_{n.pl}$]}	*

Now the two parts are realized as closely-knit syntax rather than by the juxtaposition of sentences, but the effect is similar: The object that such users are looking for is of the type *Toyota*, and they want to restrict their search to those instances that have the attribute *five doors*. This object-requesting structure consists of what we may call a **focus topic** (representing the focus feature) and a postposed **filtering modifier** (representing the modifying feature), represented in the formal notation by the optional symbol '>' pointing from the former to the latter[191]. The focus feature *constructs* the topic type whereas the modifying features merely *filter* the results. While the focus topic is a noun or noun phrase, the filtering modifier may be either adjectival or nominal. In ex. 9.21, it is a noun phrase consisting in turn of a numeral and a noun in the plural. The following examples show the use of adjectival modifiers.

9.22	**Toyota new** {T1/S2: [Toyota$_{n.sg}$] > [new$_{n.sg}$]}	*
9.23	**Toyota second hand** {T1/S2: [Toyota$_{n.sg}$] > [second hand$_{n.sg}$]}	*

In the latter example, *second hand* is basically a noun phrase, but it is used adjectivally, as in *a second hand Toyota*. However, as explained above, all adjectives not accompanying a following noun are treated as noun phrases of their own in the notation used here.

When followed by a filtering modifier, the focus topic is typically a non-specific referring expression as it tends to refer to a potential object that the user is looking for[192]. Specifically, the focus topic cannot be generic or non-referential; a request such as *TELL ME ABOUT TOYOTAS; THEY SHOULD BE SECOND HAND* simply does not make sense – you cannot modify a generic expression after it has been established. In such cases, any modification must be included in the original generic expression: *TELL ME ABOUT SECOND HAND TOYOTAS*. Indeed, this is exactly the way we tend to express ourselves if we want to know something about a topic rather than exploring objects for a purpose such as buying them. In search

[191] The symbol is optional but not redundant, since the relation is sometimes reversed, as we shall see below.
[192] Except if the modifier was added as an afterthought and hence really ought to have been preposed.

syntax, the expression is reduced to a simple noun phrase consisting of a noun with a facultative preposed **constructing modifier**.

9.24	**new Toyotas** {T1: new Toyotas$_{n.pl}$}	*
9.25	**second hand Toyotas** {T1: second hand Toyotas$_{n.pl}$}	*

As this is a construction borrowed directly from natural language, the modifier must follow the grammatical rules of the language in question. Thus, it is usually an adjective. Note, for instance, that the noun phrase *five doors* that we met as a filtering modifier would tend to be converted into its adjectival version *five-door* when used as a true constructing modifier.

9.26	**five-door Toyotas** {T1: five-door Toyotas$_{n.pl}$}	*

The hyphen may be left out in practice in queries, but the explicitly nominal plural form *five doors* would in any case not be expected to be met with in this position.

The hypothesis that the two word orders carry distinct meanings is corroborated by statistics from L^{TDC}, where 1772 people searched for *fastnet udland* ‘landline abroad’ and 460 for *mobil udland* ‘mobile (phone) abroad’, but *nobody* used the inverse constructions mentioning *udland* first[193]. This would appear to be because while *fastnet* and *mobil* are the focus topics, *udland* acts everywhere as a filtering modifier which narrows the topic down. Similarly, 2374 users searched for *internet gprs*, but not a single one chose to type *gprs internet*. Note that it cannot be said that the word *internet* is the more important in this case – quite on the contrary – but *gprs* acts as a modifier which filters the results based on the focus topic. The tendency to mention focus features first is likely to be strong and should be taken into account when retrieving results as the intentions may be exceedingly different depending on the choice of focus topic, and it is not merely a question of which topic is considered most important, either.

The recognition of focus features is complicated in English by the fact that compound nouns are not written solid. In most other languages this is not so problematic because compounds are either written solid or held together by adpositions or case forms. The phenomenon has the interesting effect that in

[193] In Danish, ’abroad’ is a noun.

English query language it may often seem as if two words may be swapped around with no difference in meaning.

9.27	**metal shelves = shelves metal**	*
9.28	**DVD shelves = shelves DVDs**	*

As indicated by the plural in the latter example, however, the syntactic function of the word *DVD(s)* differs between the two constructions. When preposed, it is the modifying part of a compound (and hence functions adjectivally), whereas when it is postposed it is an added nominal modifier that can be used in the plural. In a language such as Danish the difference is much clearer, as compounds are written solid (at least in correct use)[194].

9.29	**metalhylder = hylder metal**	*
9.30	**DVD-hylder = hylder DVD'er**	*

If a LAT succeeds in establishing that a user wants to buy a product, the system is still required to be able to decide whether the user who searched for *metal shelves* or *shelves metal* is after shelves or metal (bearing in mind that computers are unintelligent language interpreters). Simply assuming that the first word is the most important will not do.

The post-nominal modifiers used in query syntax resemble the (pre-nominal) modifying parts of nominal compounds in natural language in that they have approximately the same set of semantic relations (also known as *theta-roles*) to the head noun. This relationship means that the same intention may often lead to the use of either a natural language expression such as a nominal compound or a specifically query-adjusted construction with a post-modifier, as in ex. 9.29-9.30. A sophisticated LAT would be "aware" that *shelves metal* is synonymous with the compound lexeme *metal shelves*, and ideally what this means for the extension of the reference, i.e. that *metal* designates the material, while *DVD(s)* refers to the type of object intended to be stored on the shelves. At present, the important observation to keep in mind is that a sequence of two nouns is in no way a simple succession of terms of equal status; they are not just "sequences of keywords" (de Lima & Pedersen, 1999, p. 145). Rather, they enter into syntactic relationships which it pays to attend to.

[194] In the particular example with *DVD*, however, the two words are frequently written separately in queries because of the modifier's status as an abbreviation. Such compounds furthermore require the use of a hyphen.

9.2 Syntax and information structure

The findings of the previous section showed quite clearly that syntax is very much at work in queries, both in a form known from natural language and in a specialized form that needs to be subjected to thorough analysis (Barr, Jones, & Regelson, 2008). We cannot possibly do justice to this vast topic here, but in the following we shall present a few interesting log studies which nourish hypotheses about the reasons for the distributions observed in the data.

9.2.1 Topics and concepts

Even in the simple case of a succession of two noun phrases it is reasonable to ask how it is possible to decide which is the focus topic and which the filtering modifier. It is assumed here that by definition the first phrase is the focus topic and whatever follows is a modifier (assuming that compounds have been identified in advance and removed from this analysis as they do not contain two noun phrases); however, this assumption may require revision in future. There is obviously a possibility of adding the modifier as an afterthought – especially if the query is an extended version of a previous one without the modifier. Some of these instances might well be recoverable either from search history or from the semantics of the terms in question. Even in such cases, however, it makes sense to consider the first term to be the focus topic since that was presumably the reason for its being written first. The two terms can often be swapped around in order to achieve a subtly different nuance.

9.31	**abonnementsvilkår mobil** 'subscription terms mobile (phone)' {T1/N2: ([abonnementsvilkår$_{n.pl}$] > [mobil$_{n.sg}$])}	TDC (4)
9.32	**mobil abonnementsvilkår** 'mobile (phone) subscription terms' {T1/N2: ([mobil$_{n.sg}$]) > ([abonnementsvilkår$_{n.pl}$])}	TDC (3)

Both are here taken to represent the order Focus topic – Filtering modifier, regardless of their production history. Nevertheless, the constructions are quite different as they represent two different types of filtering; the first involves specification of a sub-concept, i.e. the concept is narrowed down from subscription terms in general to those that concern mobile phones in particular, and the other specification of a sub-topic, i.e. the topic is narrowed down from mobile phones in general to the one that only comprises the SUBSCRIPTION TERMS feature of that topic. The outcome is a concept in the former case (a set of subscription terms), and a topic in the latter (mobile phones, focusing on subscription terms). The actual information that the user is likely to get from either is very similar. However, there are some differences in likely intentions leading to these two formulations. For instance, if the user was wondering whether there are any subscription terms at all, the sub-topic version would possibly tend to be preferred, since this question is really one about mobile phones (in a given context) rather than about subscription terms. The hypothesis remains to be tested, however. After all, many other, conflicting tendencies may intervene and cause users to violate the principle, as explored more thoroughly in section 9.4 on the interaction of tendencies and countertendencies. Compare the *halogen elements* example from Cooper (1971) for further illustration of the difference (We modify the queries slightly here relative to the versions in Cooper's paper in order to disambiguate tokenization so that there is no question of where the phrase boundary is).

9.33	**chemical elements halogens** {([chemical elements$_{n.pl}$] > [halogens$_{n.pl}$])}	*
9.34	**halogens chemical elements** {([halogens$_{n.pl}$]) > ([chemical elements$_{n.pl}$])}	*

Ex. 9.33 is an instance of sub-concept filtering, from chemical elements in general to halogen elements in particular, the reason being that HALOGEN is a subordinate concept of CHEMICAL ELEMENT. The information that the user is after with this query is likely to be information on halogen elements, including but not limited to their names. Ex. 9.34, on the other hand, is a case of sub-topic filtering, from the topic of halogens to the question of what elements are included in that group. Since in the end both represent topical queries, sub-concept filtering can in fact be seen as a subtype of sub-topic filtering in which the topic is a concept. In sub-topic filtering in the narrow sense, the topics are more complex and based on domains.

It is important to note that sub-topic filtering does not involve two topics in succession as one might be led to think, but rather a topic followed by a filtering feature[195]. In other words, *chemical elements* in ex. 9.34 does not represent the topic of chemical elements but only the CHEMICAL ELEMENTS feature of the HALOGEN topic; the user is not actually interested in information on chemical elements but only in

[195] Sometimes, of course, queries do contain several topics. In such cases, the relationship between terms is much less linguistic and more dependent on typical orders of association. A study of these relationships has been started by Manshadi and Li (2009), but is still in its infancy in the IR field. They have, however, been studied for much longer in the information extraction community.

certain information on halogens. Furthermore, the user has not asked for general information on halogens but only on a particular feature. This type of query is therefore highly precise. With the sub-concept filtering method, on the other hand, the user is indeed asking for information on chemical elements, more particularly as concerns the halogens. This provides for a fuller understanding; knowing what elements are halogens in fact tells us more about chemical elements in general than it does about halogens. It reveals something about the *topic* of halogens (it specifies the CHEMICAL ELEMENT feature, as with sub-topic filtering), but only secondarily about the halogens themselves[196].

As concepts can be turned into topics, but not readily *vice versa*, the above queries could also be represented formally in the following way, using the abbreviations *F-concept* and *F-topic* for *filtering* (subordinate) concepts and topics. By using an ontology, it is possible to have a computer recognize these differences.

9.35	**chemical elements halogens** (TOPIC [CONCEPT: **chemical element**] > [F-CONCEPT: **halogen**]) {([chemical elements$_{n.pl}$] > [halogens$_{n.pl}$])}	*
9.36	**halogens chemical elements** (TOPIC [CONCEPT: **halogen**]) > (F-TOPIC [CONCEPT: **chemical element**]) {([halogens$_{n.pl}$]) > ([chemical elements$_{n.pl}$])}	*

The formal representation of queries devised here and shown at the bottom of each example is a practical simplification of this formal representation, with added information on morphology. In this concise notation, the use of round and square brackets shows what kind of filtering is employed in a given case[197]. Note that even though an F-concept is a modification of the focus topic concept, the final topic is in fact not just what is denoted by the second noun phrase; it is defined by the semantics of the filtering modifier as restricted by the sense of the focus topic.

[196] This difference may be difficult to grasp. It is much clearer in some other cases; for instance, the fact that Itzhak Perlman plays the violin does not tell us anything about violins as a concept, yet it does contribute to the *topic* of violins.

[197] Henceforth, square brackets will be omitted when the coincide with round ones, and round brackets will be omitted when square brackets are used. Thus, the appeareance of either will be enough to show the difference. Finally, round brackets may be omitted when they coincide with the curly braces.

9.37	**fremkaldelse billeder** 'development images' {[$\text{fremkaldelse}_{n.sg}$] > [$\text{billeder}_{n.pl}$]}	TDC (5)

Here, the F-concept is not just IMAGES, but IMAGE DEVELOPMENT as a subclass of DEVELOPMENT in general. This is typical of this kind of nominalizations. Clearly, these are not separate topics, but concepts that combine with each other to form a topic. It would be quite possible to turn the expression around and make an F-topic construction as in the following example. However, this happens not to occur in the TDC log.

9.38	**billeder fremkaldelse** 'images development' {($\text{billeder}_{n.pl}$) > ($\text{fremkaldelse}_{n.sg}$)}	*

It is not to be supposed that *all* instances of two consecutive terms are to be interpreted as either an F-topic or an F-concept construction. There are likely to be many exceptions. One of them is the use of discriminators of the kind used in automatic query expansion. Users also add disambiguating terms on their own accord sometimes, if they feel that it is necessary in order not to get too many irrelevant results. This is a case of what we have called *meta-awareness*. The type does not appear to be common in the logs analysed, but the following may be an example.

9.39	**thorleif musik** 'thorleif music' {[$\text{thorleif}_{n.prop}$] < [$\text{musik}_{n.sg}$]}	TDC (1)

Such disambiguation is in fact the opposite phenomenon of the F-concept construction in that the second term is broader than the first and because this superordinate concept includes the focus topic it specifies the sense of it, thereby disambiguating the term[198].

[198] There does not seem to be any sense in which the opposite of an F-topic construction might occur, since a topic can hardly be ambiguous, except as regards certain terms used in reference to it, which would result in a construction of the type just illustrated.

9.2.2 Evidence of compliance with the principles

As a result of the different semantics of the F-concept and F-topic constructions we would not expect users to issue the query *computer new* when looking for a new computer to buy. After all, this would be paraphrasable as *I WOULD LIKE TO BUY A COMPUTER; IT SHOULD BE NEW*, and this is not usually the intended sense of such a query (unless interpreted as being in contrast to second-hand computers). This supposition is corroborated by L^{TDC} where we find 30 people searching for *ny computer* but none with the opposite word order.

9.40	**ny computer** {ny $\text{computer}_{n.sg}$}	'new computer'	TDC (30)

The same effect is seen with the adjectives *flere* 'more' and *ekstra* 'extra', which are likewise invariably preposed, for similar reasons. On the other hand, most other adjectives are predicted to be able to occur in both positions, with a corresponding difference in intention, and this is borne out by the data.

9.41	**langsom pc** {S1: langsom $\text{pc}_{n.sg}$}	'slow pc'	TDC (259)
9.42	**pc langsom** {S1: ($\text{pc}_{n.sg}$) ($\text{langsom}_{n.sg}$)}	'pc slow'	TDC (11)
9.43	**gratis ringetoner** {N2: gratis $\text{ringetoner}_{n.pl}$}	'free ringtones'	TDC (111)
9.44	**ringetoner gratis** {N2: [$\text{ringetoner}_{n.pl}$] [$\text{gratis}_{n.sg}$]}	'ringtones free'	TDC (22)
9.45	**hvid iphone** {S2: hvid $\text{iphone}_{n.sg}$}	'white iphone'	TDC (48)
9.46	**iphone hvid** {S2: [$\text{iphone}_{n.sg}$] [$\text{hvid}_{n.sg}$]}	'iphone white'	TDC (40)

Fig. 9.5 shows some statistics regarding the use of adjective + noun vs. noun + adjective orders, and also of two nouns in succession. Only queries that contain nothing apart from these constellations were included in this analysis.

	All searches	Distinct searches
A + N	31,795	380
N + A	3,093	47
N + N	197,610	2,597

Fig. 9.5. Statistics regarding the use of adjective-noun vs. noun-adjective and noun-noun constellations in L^{TDC}. These numbers were calculated by automatic analysis and hence may contain a small amount of errors due to misanalysis, such as *ring hemmeligt* 'call secretly' where the imperative *ring* 'call' was misinterpreted as the homophonous noun 'ring', but such errors seem to be few in number so the results are deemed to be reliable.

As can be seen from the table, the simpler structure with a preposed adjective and hence only one noun phrase is more than ten times as frequent as the opposite order which represents two noun phrases according to our definition. However, the above examples show clearly that this is highly variable across actual cases; some combinations of adjective and noun are much more likely to occur frequently in both constructions than others. Thus, with SLOW PC 96% of searches use the adjective-noun order, whereas with FREE RINGTONES 83% do, and with WHITE IPHONE only 55%. While the explanation of this substantial difference is not immediately self-evident, it is only natural to suppose that there might be a reason behind it, and a closer look reveals that this seems indeed to be so. Ex. 9.41-9.42 describe a problem that needs solving. What is important about the problem here is not the pc but the fact that it is slow, and consequently the more marked construction with a postposed adjective is dispreferred. In order to understand the subconscious reasoning behind the other differences, we shall need to look closer at resultative constructions.

9.2.3 Syntactic search styles

As noted earlier, adjectives and participles may be postposed if they are resultative, in order to focus on an event and its consequences.

9.47	**mobil tabt** 'mobile (phone) dropped' {S1: $mobil_{n.sg}$ $tabt_{res}$}	TDC (6)

9.48	**tabt mobil** 'lost mobile (phone)' {S1: tabt mobil$_{n.sg}$}	TDC (128)

Ex. 9.47 is resultative; it focuses on the event that has passed. As such, it is most likely a query as to the consequences of dropping a mobile phone. Ex. 9.48 simply describes the problematic state that the user is in through its currently most prominent and relevant characteristic. Hence, this query is really more about the user than about the phone; the user is likely to be wondering what to do next. The resultative construction appears to be in frequent use in order to indicate that the state is a problematic one, whereas the opposite order is neutral as to whether the state described is problematic or desirable.

9.49	**mobil stjålet** **telefon stjålet** **mobiltelefon stjålet** 'mobile/phone stolen' {S1: mobil$_{n.sg}$ stjålet$_{res}$}	TDC (96) TDC (32) TDC (16)
9.50	**mailbox fuld** **mailboks fuld** **postkasse fuld** 'mailbox full' {S1: mailbox$_{n.sg}$ fuld$_{res}$}	TDC (44) TDC (36) TDC (20)
9.51	**simkort låst** **sim kort låst** 'sim card locked' {S1: simkort$_{n.sg}$ låst$_{res}$}	TDC (28) TDC (13)

9.52	**simkort afvist** 'sim card rejected' {S1: simkort$_{n.sg}$ afvist$_{res}$}	TDC (13)
9.53	**mobil spærret** 'mobile (phone) blocked' {S1: mobil$_{n.sg}$ spærret$_{res}$}	TDC (6)

Perhaps it would be better to say that the construction is ambiguous, given that this ambiguity is resolved when a determiner such as a possessive pronoun is inserted between participle and head, thus showing beyond any doubt that it is indeed a resultative construction.

9.54	**tabt min mobil** 'lost my mobile (phone)' {S1: tabt$_{v.res}$ min mobil$_{n.sg(1p)}$}	TDC (3)

Even the query *pc langsom* 'pc slow' in ex. 9.42 belongs marginally to the resultative type. However, it is not prototypical in that it probably does not describe anything like a sudden change of state but simply an accumulating annoyance. If this is the case then the construction is not a true resultative in this instance, which explains its low frequency. The (infrequent) occurrence of this construction may be ascribed to natural random variation and/or analogy with other problem-describing queries. Ex. 9.43 and 9.45 are both cases of wanting to buy or download a product. These are straightforward cases of sub-concept filtering from a certain class of product to a specific sub-type. Hence, there is indeed a reason for using the postposed adjective in these examples. Nevertheless, the construction would not be expected to be quite as frequent as the one with a preposed adjective because it is more marked (containing as it does two noun phrases rather than one), and this is again borne out by the data. The reason that *iphone hvid* is so much more common relative to *ringetoner gratis* may have several related possible explanations, which may all have contributed to the surface distribution. Whereas WHITE is a true feature of the product, FREE is part of the very specification of what the user needs, and *gratis ringetoner* is basically a topical query, or at least structured on the pattern of one with a generic topic, which the reader may recall does not allow post-modification. The query is hence closely related to the problem-describing type; more particularly, it is a type of query in which the user tells the system what he or she needs in the current situation instead of describing the object which might potentially satisfy that need, as is the case with the query *hvid iphone*. In terms of the Stage-based Intent model, the three requests accordingly represent the Source-, Aim- and Goal-targeting types, respectively. Interestingly, the log data hints at a possible correlation between these types and the preferred syntax in such queries. The exact inventory of mapping rules or principles would have to be identified in future, but at least the examples seem to indicate that such mappings do exist[199]. A likely explanation is that, being situations, Source and Goal are targeted with descriptive phrases, including filtering modification, whereas other targets normally involve specification and classification rather than description and therefore do not allow descriptive modification but only restrictive qualifiers, which tend to be preposed.

In order to ease discussion of these structures and behaviours, we may call the version with a preposed constructing modifier the **constructing query style (CqS)** and the type with a postposed filtering modifier the **filtering query style (FqS)**, except when the modifier is resultative, in which case we shall call it the **resultative query style (RqS)**. Note that the latter two constructions are not always distinguishable on the surface. The use of these three styles is not entirely straightforward, but a few suggestions as to their preferred usage may be made. We shall not attempt to prove these hypotheses at present as that would require us to go beyond the purpose of this thesis but merely present them as suggestions providing food for thought.

- CqS is probably to be considered the more basic type; it appears to be the default choice when searching for information on a topic.
- FqS seems to be the default choice when looking for objects (e.g. wares), individuals, files (e.g. images, music, videos), etc.
- CqS may probably always be chosen where FqS would be expected, because the user may opt to search for information on the object type rather than for the object itself.

[199] Note that the syntactic descriptions attempted here should only be taken to hold for English. In a language such as French, where modifiers are usually postposed, the distinction can hardly be upheld in the same way as in English or Danish. It would be highly interesting to study whether a different way of upholding the distinction has developed among French users. That is, however, beyond the scope of the present thesis.

- RqS seems to be the default choice when searching for information on how to cope with objects that fulfil certain roles, such as how to treat an object which is already in the user's possession.

The last principle takes care of the resultative construction. Here is another example.

9.55	**Toyota rusty** {[Toyota$_{n.sg}$] > [rusty$_{n.sg}$]}	*

Probably few would consider searching for *rusty Toyotas* in order to be told how to deal with the rust, because the reference is to a specific Toyota which has become rusty and not to the generic set of rusty Toyotas. The CqS version would rather seem to imply that the user wanted a rusty Toyota than one without rust. The same holds for the example *iphone låst* 'iphone locked', which in its FqS form implies a resultative problematic state rather than a description of the iPhone as having the characteristic that it is locked.

This may seem contradictory; why is CqS suddenly more indicative of a wish to buy the object than FqS? The answer is that it isn't. The FqS version is just as suitable for the purpose of buying rusty Toyotas, only the interpretation in which the user wants information on how to cope with the rust is far more likely[200]. The CqS version, on the other hand, is basically for finding information, but can be used as a first step in buying something, and this interpretation may even seem marginally more likely than the one where the user is simply highly interested in information on rusty Toyotas. Thus, the intentions seem reversed, but they are basically constant. Obviously, as the use of query syntax is mostly subconscious and not formalized, there is no guarantee that a particular user follows the principles as stated here, but it is predicted that there will be a statistical tendency towards compliance. A study of actual user behaviour in this regard is left for future work. It may be that future studies will show that I have been too confident in users' reliance on linguistic structures that make sense and that queries are indeed to a large extent merely *word salads*. However, unless otherwise proven I do think that the linguistic structures discussed above are so basic and deeply rooted that people cannot but follow such principles, even though a certain amount of arbitrariness will always be present. Also, different syntaxes may have evolved in different user societies so that we cannot be sure that everyone follows the same rules. This is an interesting field worth much further investigation.

[200] Note that the alternative version *rust Toyotas*, in which the rustiness has been nominalized, most likely treats *rust* as the focus topic and *Toyotas* as a modifier.

9.2.4 Syntax and pragmatics

From a general linguistic point of view, it is highly interesting that there seem to be two kinds of syntax in search; **semantic syntax** and **pragmatic syntax**, as it were. Traditional linguistics treats such notions as impossible, because syntax is normally seen as complementary to both semantics and pragmatics. However, it would certainly seem that some kinds of syntax are used as part of the *informative intention* and others as part of the *communicative intention* (see section 4.4). The former consist of syntactical constructions that also occur in natural language English, whereas the latter are probably all specific to search; these are filtering expressions that specify the search further in a much more "algorithmic" way than what is normally allowed in NLE. It is accordingly claimed here that there is a difference not only in form but in kind between the following two queries.

9.56	**used car**	*
9.57	**car used**	*

The former is an example of a modifier construction borrowed directly from NLE, whereas the latter (which is not resultative in the default interpretation) is a filtering modifier construction that is specific to search. The focus topic in ex. 9.76 is *car*, but with the further specification that only the used ones interest the user. So far, we have treated both of these constructions as syntax, but added that the latter kind follows rules that do not hold in NLE. However, the differences run deeper than that; it can be argued that the latter is not an instance of syntax as generally understood but only in a broader sense which includes some pragmatic phenomena. Recall that the two terms are here considered to consist of two co-referential noun phrases and not one, which makes it similar to the use of appositional nouns in natural language. The construction functions interpersonally as it were (in spite of the fact that the interlocutor is a machine), corresponding to the paraphrase *I'M LOOKING FOR A CAR; IT SHOULD BE USED*, in which the second clause is more of a meta-level instruction to the hearer than a semantic modification of the focus topic. This is a kind of "pragmatic syntax" which is not there to help in the construction of semantic descriptions as other syntax but as a means to getting an intention across.

Appositions may in fact likewise be considered a case of what we have called pragmatic syntax. The same could be said of afterthoughts and preposed topics (and postposed ones, for that matter), as in the following examples from NLE (Lambrecht, 1994, p. 193)[201].

[201] Such topics are much more common and more strictly grammaticalized in many Asian languages in particular.

9.58	The typical family today, the husband and the wife both work.
9.59	The African elephant, it's so hot there, so he can fan himself.

Not being a part of the main sentence, the excised topic is not really part of the informative intention conveyed by it either. Rather, it is an addition that makes the information useful in conveying the communicative intention by attracting the hearer's attention to certain topics and embedding the sentence in the discourse context.

The perceptive reader will have noticed that "pragmatic syntax" is an important component of what is known as **information structure**. In natural language that notion may well be sufficient, but in search it appears rather to be a kind of syntax on the pragmatic level related to the communicative intention. This is because the recipient is a machine and is consequently entirely dependent on form; to a LAT, the difference between CqS and FqS is merely one of word order, and consequently it would be difficult to argue that just because the word order is different it is no longer a question of syntax. Consequently, it must be considered a special kind of syntax which carries out pragmatic functions.

9.3 Other variation and behavioural phenomena

As has been shown, there are intricate relationships between the words in a query, and the order is in no way fortuitous, despite the fact that words may indeed be added as an afterthought and rules may be broken because people do not necessarily plan ahead before typing. The fact that such rules are frequently broken should not discourage the analyst. Instead, more sophisticated rules will have to be developed in future regarding various uses and situations. The important point to note here is the fact that these phenomena occur at all. In-depth studies should be made of the choices made by searchers and the reasons for such choices, in order that we may understand the way intentions and expressions map onto each other. A small contribution to this important discussion is given below, but a thorough treatment is not within the scope of the present thesis.

The various syntactic and morphological phenomena that we have found interact in interesting ways. This is clear from the following study of queries that combine the name of a mobile brand with a form of one of the words *mobil* and *mobiltelefon* in L^{TDC}.

9.60	**samsung mobil** 'samsung mobile phone' {T1/T2: samsung mobil$_{n.sg}$} / {T1/T2: (samsung$_{n.prop}$) (mobil$_{n.sg}$)}	TDC (50)
9.61	**samsung mobiler** 'samsung mobile phones' {T1/T2: samsung mobiler$_{n.pl}$} / {T1/T2: (samsung$_{n.prop}$) (mobiler$_{n.pl}$)}	TDC (11)
9.62	**mobil samsung** 'mobile phone samsung' {T1/T2: (mobil$_{n.sg}$) (samsung$_{n.prop}$)}	TDC (7)
9.63	**mobiler samsung** 'mobile phones samsung' {T1/T2: (mobiler$_{n.pl}$) (samsung$_{n.prop}$)}	*

Fig. 9.6a-b present the relevant statistics. It appears that mentioning the brand first is more common than the opposite, and also that singular expressions are more frequent than plurals.

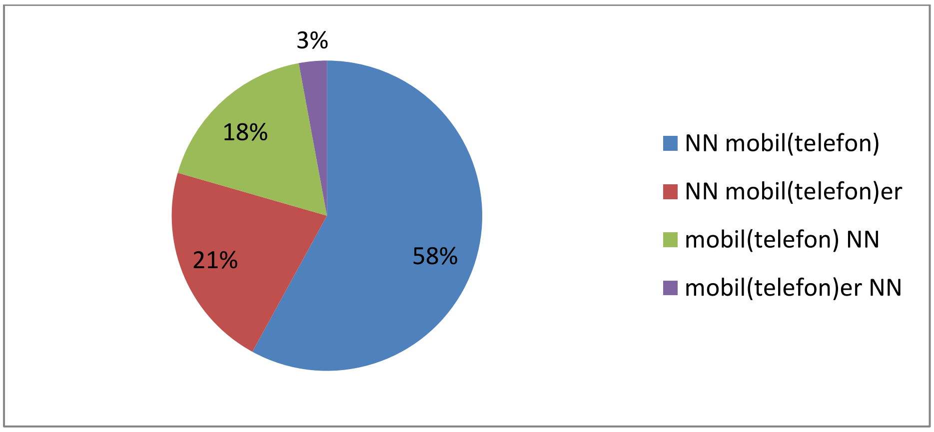


Fig. 9.6a. Statistics on expressions involving the name of a mobile phone producer and a form of one of the synonyms *mobil* and *mobiltelefon* in L^{TDC}. The forms ending in *–er* are plurals.

	Singular	Plural
NN mobil(telefon)	59,6 % (618)	21,6 % (229)
mobil(telefon) NN	16,0 % (188)	2,8 % (31)

Fig. 9.6b. Statistics on expressions involving the name of a mobile phone producer and a form of one of the synonyms *mobil* and *mobiltelefon* in L^{TDC}.

When pre-posed, the brand name is usually a modifier in a tightly-knit phrase in the manner of the pre-posed adjective in a phrase such as *four-door Toyotas*[202]. With the brand name postposed, on the other hand, the effect is similar to *Toyotas four doors* – there are now two phrases of which the second has a filtering function. It is especially striking that the construction *MOBIL(TELEFON)ER NN* is so rare. It appears that the intention that is supposedly correlated with the use of plurals usually demands the mention of the producer before the general term describing the product. This is one case where linguistics is able to provide an explanation. The plurals are clearly generic expressions, i.e. they refer in principle to all instances of a given type. People who search for *mobile phones* probably want to see *all* (relevant) mobile phones which are available or can be found. However, as already mentioned, a generic expression cannot be assigned a restrictive modifier because it refers to all instances of a type by definition. In other words, you cannot ask for all mobile phones and then add that they (i.e. all mobile phones) should have been made by Nokia. Conversely, it is entirely possible to modify the concept MOBILE PHONES itself before turning it into a generic expression. Thus, the expression *Nokia mobile phones* is *one* generic noun phrase where *Nokia* modifies the root sense of the concept MOBILE PHONE before the entire thing is turned into a generic phrase[203]. This is a straightforward case of generic reference. The phrase *mobiltelefoner Samsung* is not impossible, it is just that it must be interpreted as an abbreviation of a fuller phrase such as *mobiltelefoner af mærket Samsung* ‘mobile phones of the brand Samsung’. There is no explicitly expressed relation between the terms in such a query, which makes it less transparent both for the human reader and for the system. This may contribute to the avoidance of such constructions.

In the following, we shall explore some other highly interesting kinds of expressive variation in queries. The first is lexical variation, representing the scale and qualitative aspect chosen, followed by functional variation, such as the use of questions and imperatives. Next, we shall be looking at stadial variation in the sense of target choice, and lastly the way users refer to actions they want performed.

[202] Note that the brand name corresponds to the adjective, while the noun corresponds to *Toyotas*.
[203] Alternatively, it is possible to interpret this construction as consisting of two phrases, but nothing seems to be won from such an interpretation.

9.3.1 Lexical variation

Compared to morphological and syntactic variation, lexical choice in reference to the same kind of entity is even harder to predict and exploit. For example, users who are interested in *toyotas* may in the end settle on searching for *cars*, and *vice versa*. The reasoning behind this is complex and involves both the extent to which they have in fact decided on a make of car and whether they think it likely that issuing the query in question in the given circumstances will result in a useful response. The choice may also be influenced by the user's anticipation of the site structure, as when a user searches for *cars* in order to be taken to the correct page on a site (as opposed to the *motorcycle* page, for instance). This may then be followed by navigation and/or search in order to specify the car make and model[204].

One may easily be deceived into thinking that each word in our lexicon corresponds to a specific ontological class and that consequently a speaker with a certain type of entity in mind has simply to choose the correct term in order to refer successfully to that type of entity, but we actually have a variety of words at our disposal no matter what kind of entity we want to refer to. The choice depends partly on context and partly on the **scale** and **qualitative aspect** of the entity that we mean to focus on, as discussed in section 5.3. For instance, a statue of Venus may be given any of the following denominations: *A statue*, *Venus*, *a piece of art*, *a lump of clay*, *a ruin*, or *a battering ram*, along with an infinite number of others of varying complexity. These are aspectual versions of the same concept (Masolo, Borgo, Gangemi, Guarino, & Oltramari, 2003). This particular example may be thought extreme, but aspectual variation is a common problem in search ontologies. Just consider the case of *rats*, which some may refer to as *pets* and others as *pests*; the choice is hardly insignificant. Even though these terms refer to exactly the same animal, resources on rats-as-pets and rats-as-pests are not relevant to the same people or in the same situations, and thus should be kept as clearly apart as texts on love and war. Furthermore, different levels of generality may be assumed, as in the choice of any of the terms *animal*, *pet*, *rodent*, *rat* or *Black Rat*. There are thus two main sources of variation in the choice of term: 1) speakers focus on different aspects of the same entity, and 2) speakers choose different levels of generality, i.e. different hierarchical levels of terms.

As already pointed out, lexical variation is difficult to study, and we shall once more content ourselves with demonstrating that it does indeed occur. The following small-scale study, which is an extension of the one referred to above, exhibits several of the variational phenomena in question (fig. 9.7).

[204] This is a kind of *orienteering behavior* as discussed in Teevan et al. (2004).

	Form	Samsung	%	Sony	%	Nokia	%	HTC	%
mobiltelefon NN	indef.sg.	4	2,8	4	2,2	20	4,3	3	1,1
mobiltelefoner NN	indef.pl.	0	0,0	4	2,2	8	1,7	0	0,0
NN mobiltelefon	indef.sg.	28	19,6	24	12,9	48	10,4	12	4,3
NN mobiltelefoner	indef.pl.	4	2,8	5	2,7	35	7,6	11	3,9
mobil NN	indef.sg.	16	11,2	19	10,2	82	17,8	40	14,3
mobil, NN	indef.sg.	1	0,7	0	0,0	1	0,2	0	0,0
mobilen NN	def.sg.	0	0,0	0	0,0	1	0,2	0	0,0
mobiler NN	indef.pl.	0	0,0	5	2,7	3	0,7	11	3,9
NN mobil	indef.sg.	78	54,6	71	38,2	214	46,3	143	51,1
NN, mobil	indef.sg.	0	0,0	0	0,0	1	0,2	0	0,0
NN mobilen	def.sg.	0	0,0	0	0,0	0	0,0	1	0,4
NN mobiler	indef.pl.	12	8,4	54	29,0	49	10,6	59	21,1
		143		**186**		**462**		**280**	

Fig. 9.7. The distribution of forms of the words *mobil* and *mobiltelefon* in L^{TDC}. NN represents any of the telephone brands included in the study.

It was performed on the subsection of the TDC log containing the interchangeable words *mobil* 'mobile (phone)' and *mobiltelefon* 'mobile phone' in combination with a selection of telephone brands (*Samsung*, *Sony*, *Nokia*, *HTC*). The table can be approached from a morphological, syntactic or lexical viewpoint, depending on the phenomenon that one wishes to investigate. The study reveals a lexical choice between the two alternative terms, and also shows variation in morphological form (singular vs. plural in particular), and in word order. It provides a chance to see what may happen with lexical choice given a diverse set of users, because when the choice between the terms *mobil* and *mobiltelefon*, both meaning 'mobile phone', is correlated with the brand of mobile phone that users are looking for, an interesting distribution is revealed (fig. 9.8). Apart from the simple observation that the abbreviated term is more than four times as common as the full version, it seems that the long – and slightly more old-fashioned – term is noticeably less often utilized by people interested in HTC products than with the other brands analysed. This could be taken as an indication that there are indeed different user groups with differing linguistic preferences, which in this case may even be correlated to other preferences, such as preferred mobile phone brands. Obviously, categorizing users according to their preferences as regards mobile phone brands would be somewhat *ad hoc*, but it is nonetheless interesting that such variation should appear so clearly from the log material. One should be prepared to come across this kind of variation everywhere in search logs, though it can often be hard to spot because you have to be looking for it first in order to notice it.

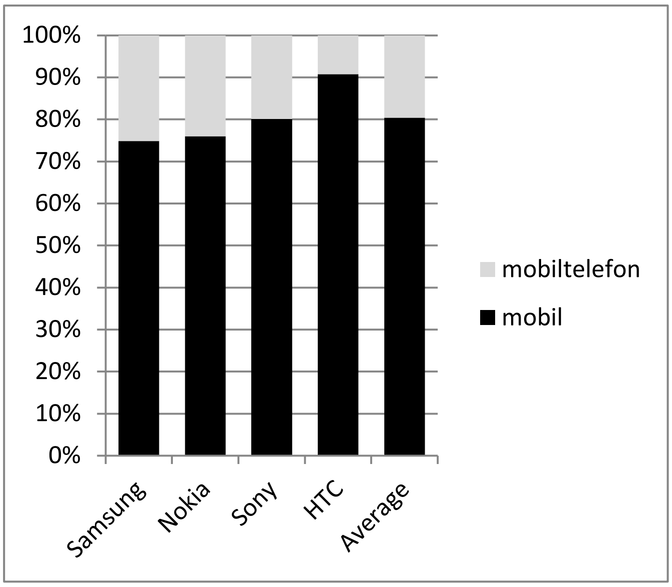


Fig. 9.8. The choice between the lexemes *mobil* and *mobiltelefon* for the concept MOBILE PHONE in L^{TDC} reveals a difference between sets of users looking for different brands.

Studying lexical variation in queries in general is difficult; the best we can do at present is provide an explanatory theory on which to build a better understanding of word choice. Hopefully, the matter will receive due attention in future. The above small-scale study provides a mere glimpse into the extensive lexical variation that exists between users. The result indicates that the potential uses of lexical choice identification differ from those regarding syntax and morphology: Rather than exploiting it for the purpose of determining user intention, lexical choice may be employed for recommendations based on similar users. In addition to this, the chosen scale tells us something about the user's specificity of decision; searching for *cars* rather than *Toyotas* may signal a greater openness towards alternative suggestions.

9.3.2 Functional variation and verb morphology

We shall include under the heading of *functional variation* the use of interrogative expressions and imperatives, as opposed to other queries with a more "neutral" function (however that function is interpreted). Both interrogatives and imperatives signal special kinds of queries which tend to reflect special intentions. In this section, we shall study functional variation by looking at the use of verbs in L^{TDC}. In order to ensure similar contexts across instances we shall restrict our attention to verbs that *occur initially in a query with a frequency above 10*. Present forms, for instance, may often occur in the middle of whole sentences in queries; in such cases, the entire sentence has been borrowed from natural language, and as such the actual form of the verb is of less interest, though the fact that a whole sentence is used is in itself highly significant. When a present is used at the beginning of a query, however, it does not normally represent an entire sentence, and so special query phenomena must be at play. The exception is verb-first interrogative sentences, which are highly restricted in their occurrence in English but very common in Danish and so had to be removed by hand after the initial dataset had been generated. Since questions to TDC tend to revolve around the same range of actions and hence to be restricted to a few verbs, this was not as laborious as it might otherwise have been (initial verbs tended to be copulas and auxiliary and modal verbs). Appendix D2 lists the resultant dataset of verbs along with the frequency of the various forms in which they occur.

Many verbs can be used in several different senses or subsenses, and thus whenever a context was present which allowed disambiguation (typically in the form of a following noun phrase or prepositional phrase) the verb was assigned to one of a set of different senses defined by such contextual clues (fig. 9.9). Similar contextual phrases were grouped together.

Contextualized verb		Imperative use
oprette (service)	'set up (service)'	100%
oprette (bruger)	'create (user)'	81%
oprette (mailadresse)	'create (e-mail address)	75%
oprette (hjemmeside)	'create (homepage)'	65%

Fig. 9.9. Context-derived senses of the verb *oprette* occurring in L^{TDC}, along with the percentage of the occurring instances that are used in the imperative. It is clear that different subsenses are not equally frequent in the imperative, and this might very well reflect an identifiable principle related to the semantics of the verb and (especially) the situations in which they are used in a query; i.e. certain types of problems or needs may cause the user to prefer an imperative more often than others.

Whenever a verb occurring in different senses could not be assigned to any specific sense due to a lack of contextual clues it was given the empty contextual description ().

The reason that this partitioning of verb senses was that it seemed likely that different senses of the same word might occasion different behaviour as regards the use of various forms. It is clear from the data in Appendix D2 that verbs (and different senses of verbs) differ in the frequency with which they are used in the various forms, and it is conceivable that this difference in behaviour might reflect a deeper difference in intentions which could be exploited for decoding intents. This is what is attempted in section 9.4 where the verb senses are kept apart for this reason.

The number of unambiguous imperatives is very great, at 42,165 queries distributed over 98 different verbs, compared to a mere 15,979 unambiguous infinitives distributed over 140 verbs. Furthermore, 0.3% of all searches in L^{TDC} were questions of various kinds[205]. Appendix D3 displays the relevant data. The question is of course what intentional differences this choice between interrogative, imperative and "neutral" queries represents. We shall find that it does indeed represent a very important factor, which is closely related to the next type of variation, *stadial variation*, as the use of imperatives tends to correspond to the Aim target whereas infinitives and nominalizations appear to be preferred with Subject targeting-queries.

9.3.3 Stadial variation

It is a basic assumption in much of the IR literature that a query represents the user's needs, and this is often taken to mean that it refers to what the user wants to find. While this may certainly be so in many cases, a substantial proportion of searches do not in fact refer to the thing sought but rather describe the problem the user is facing, or something entirely different. **Stadial variation** in search is a highly interesting study, not just because it is so obviously connected with what the user is looking for, but also because of the possibility that the user may not be referring to the desired object or even a topic at all, but to something else. This is especially common in some domains where improving the personal circumstances of the user or of something in the user's possession is the Goal of the search, such as problem solving or job search. For a few illustrative examples, consider the following queries from L^{TDC}.

9.64	**utilfreds kunde** 'dissatisfied customer'	TDC (2)
9.65	**kommende kunde** 'future customer'	TDC (1)
9.66	**ny kunde** 'new customer'	TDC (132)

The three occurrences of the word *kunde* 'customer' in these examples do not refer to the same thing. In ex. 9.64, it refers to the unsatisfied user – the problematic state as it were, while in ex. 9.65 it refers to a desired state that the user wants to enter. The future customer is of course still supposed to be the user, but the "question" is not *WHAT TO DO, GIVEN THAT I AM A FUTURE CUSTOMER*, but rather *HOW DO I BECOME A (FUTURE) CUSTOMER*. In ex. 9.66, the word can have either of these representations, i.e. the

[205] Simple questions without the use of question marks were excluded from the analysis for reasons of simplicity (they are hard to locate automatically). They are not likely to be very common.

query may be paraphrased as either *WHAT TO DO, GIVEN THAT I AM A NEW CUSTOMER*, or *HOW DO I BECOME A NEW CUSTOMER*. The term is thus ambiguous in its intended sense[206], even though the lexeme *kunde* does not in itself present any ambiguity. This kind of variation is quite hard to detect and exploit, because you cannot know for certain what the user is referring to. It may, however, be of no slight importance. Consider the following queries from the Workindenmark job portal.

9.67	**engineer copenhagen**	WID (24)
9.68	**phd biology**	WID (21)
9.69	**cad copenhagen**	WID (10)

Does ex. 9.67 imply that these users wish to work as engineers in Copenhagen, or that they are engineers living in Copenhagen and looking for a job in the vicinity? We simply cannot tell. Does ex. 9.69 imply that these users have experience with cad or that they want to work with it but do not necessarily have any experience? In this case, it is perhaps most likely that they are referring to existing skills, but at the same time that they are looking for a job in Copenhagen, i.e. the first term *cad* refers to the pre-search state, *Stage 1*, while the second, *copenhagen*, refers to a feature of the desired state, *Stage 2*. The query *dissatisfied customer* can hardly be mistaken for a reference to a desired Stage 2, but then computers are notoriously unintelligent. It is vital that we enable systems to make an attempt at distinguishing between stages. The entire Stage-based Intent model suggested in chapter 8 is built on this basic referential distinction, which as we have seen is much more complex than these simple examples suggest.

9.3.4 Solving a problem

It would be interesting at this point to study the way users actually go about solving their problems using a search engine. For this purpose, a small-scale study was made of the 1973 separate searches in L^{TDC} which refer directly to problems that need solving (i.e., they include a form of the word *problem*) (fig. 9.10).

[206] A note on intention: The *intended sense* is the one in which a term must be interpreted in order to satisfy the user's needs that caused him or her to phrase the query in a particular way. It does not follow that the user is necessarily aware of this intention.

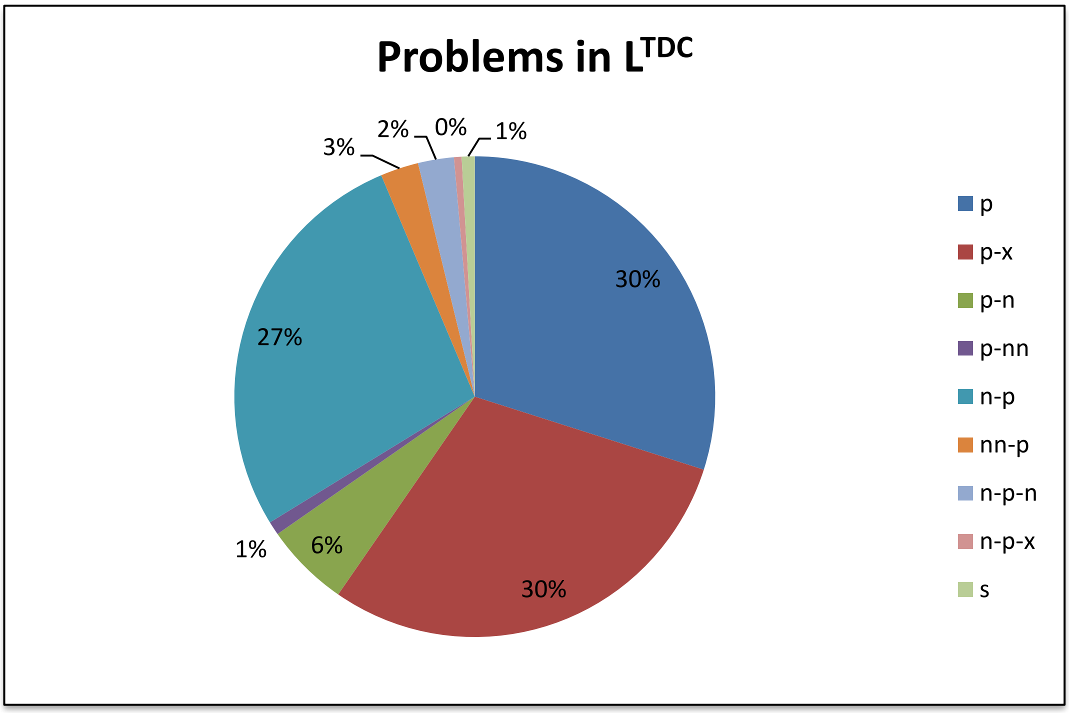


Fig. 9.10. Distribution of problem-reporting constructions in L^{TDC}. The signature *p* represents the word *problem* and compounds containing it; *x* represents a prepositional phrase added as in natural language (*problemer med*… 'problems with…'); *n* a topic noun phrase; *nn* a complex topic (two or more intersecting topics); *s* an entire sentence describing the problem.

It is clear from this analysis that there is great variability in the way people express the fact that they have problems they need help with. 583 of the queries only mention the problem, i.e. they consist solely of the word *problem* or a compound containing it as the last member.

9.70	**emailproblemer** 'e-mail problems' {S1: emailproblemer$_{n.pl}$}	TDC (1)

Another 579 searches add a prepositional phrase which describes the problem, in a construction borrowed from natural language.

9.71	**problemer med at sende mail** 'problems with sending mail' {S1: problemer$_{n.pl}$ med$_{prep}$ at_sende$_{v.inf}$ mail$_{n.sg}$}	TDC (15)

111 searches add a topic without the preposition.

9.72	**problemer mail** 'problems mail' {S1: (problemer$_{n.pl}$) > (mail$_{n.sg}$)}	TDC (4)

In 18 cases, the topic is complex and consists of two or more topics, of which one typically represents what the trouble is and the other what it concerns.

9.73	**problemer mail delay** 'problems mail delay' {S1: (problemer$_{n.pl}$) > (mail$_{n.sg}$) > (delay$_{n.sg}$)}	TDC (1)

One of the topics may be in the form of a prepositional phrase.

9.74	**problemer TDC play på ipad** 'problems TDC play on ipad' {S1: (problemer$_{n.pl}$) (TDC_play$_{n.prop}$) (på$_{prep}$ ipad$_{n.sg}$)}	TDC (2)

In 534 searches, the topic is pre-posed rather than post-posed.

9.75	**mail problemer** 'e-mail problems' {S1: (mail$_{n.sg}$) > (problemer$_{n.pl}$)} / {S1: (mail problemer$_{n.pl}$)}	TDC (63)

This type is in fact often ambiguous, because it may alternatively be interpreted as a compound noun which is – incorrectly – not written solid. In a few instances this is clearly the case, because the first noun has a special form reserved to compound use. Such cases were counted as compounds.

9.76	**hastigheds problemer** 'speed- problems' {S1: (hastigheds problemer$_{n.pl}$)}	TDC (1)

50 queries have a complex preposed topic.

9.77	**iPhone SMS problem** {S1: (iPhone$_{\text{n.sg}}$) > (sms$_{\text{n.sg}}$) > (problem$_{\text{n.sg}}$)}	TDC (3)

A complex topic may in fact be distributed before and after the *problem*-word. This happens in 47 cases[207].

9.78	**Mail problemer iPad** 'Mail problems iPad' {S1: (Mail$_{\text{n.sg}}$) > (problemer$_{\text{n.pl}}$) < (iPad$_{\text{n.sg}}$)}	TDC (1)
9.79	**mail problemer password** 'mail problems password' {S1: (mail$_{\text{n.sg}}$) > (problemer$_{\text{n.pl}}$) > (password$_{\text{n.sg}}$)}	TDC (1)

In most cases, the preposed topic and the problem word may once more be interpreted as a compound, but in a few instances this is not possible.

9.80	**sikkerhedspakken problem adgang til internet** 'the safety package problem access to internet' {S1: (sikkerhedspakken$_{\text{n.sg.def}}$) > (problem$_{\text{n.sg}}$) > (adgang$_{\text{n.sg}}$ til$_{\text{prep}}$ internet$_{\text{n.sg}}$)}	TDC (1)

In 10 searches, the postposed topic is expressed in a prepositional phrase.

9.81	**Sikkerhedspakke Problemer med opdatering** 'Safety package Problems with update' {S1: (Sikkerhedspakke$_{\text{n.sg}}$) > (Problemer$_{\text{n.pl}}$ med$_{\text{prep}}$ opdatering$_{\text{v.nom}}$)} / {S1: (Sikkerhedspakke$_{\text{n.sg}}$) > (Problemer$_{\text{n.pl}}$ med$_{\text{prep}}$ opdatering$_{\text{v.sg}}$)}	TDC (1)

Finally, 17 searches contain entire sentences that describe the problem state (Stage 1).

[207] Note the difference in analysis of the two examples shown.

9.82	**jeg har problemer med at sende e-mails** 'I have trouble sending e-mails' {S1: jeg_{pron} $\text{har}_{\text{v.pres}}$ $\text{problemer}_{\text{n.pl}}$ med_{prep} at $\text{sende}_{\text{v.inf}}$ $\text{e-mails}_{\text{n.pl}}$}	TDC (5)

It seems that either the topic is incorporated in the problem-signalling word so as to form a compound (30%), or a construction from natural language is used (*problemer med* 'problems with', etc.) (30%); alternatively, the problem topic is mentioned first, followed by the problem-signalling word (27%). Few searches place the topic after the problem-signalling word without using a natural language expression. This indicates that there is a standard query syntax which demands that the topic be stated first, but which allows natural language expressions to be inserted in their entirety.

There are also some cases where the user aims for a solution tool rather than the actual solution.

9.83	**support telefonnummer** 'support phone number' {S2: ($\text{support}_{\text{n.sg}}$) ($\text{telefonnummer}_{\text{n.sg}}$)}	TDC (1)

In the TDC-log, 16 searches are of this kind.

9.84	**windows 7 problemløsning** 'windows 7 troubleshooting' {T1: $\text{windows\_7}_{\text{n.prop}}$} {N2: $\text{troubleshooting}_{\text{v.nom}}$} / {T1/N2: ($\text{windows\_7}_{\text{n.prop}}$) ($\text{troubleshooting}_{\text{v.nom}}$)}	TDC (1)
9.85	**hjælp til pc-problemer** 'help with pc problems' {N1: $\text{hjælp}_{\text{v.nom}}$ til_{prep} $\text{pc-problemer}_{\text{n.sg}}$}	TDC (1)
9.86	**problemløsning** 'problem solving (troubleshooting)' {T1/N2: $\text{problemløsning}_{\text{v.nom}}$}	TDC (1)

9.87	**telefonnommer angående problemer med internet** '[phone number] regarding problems with internet' {N2: $telefonnommer_{n.sg}$ $angående_{prep}$ $problemer_{n.pl}$ med_{prep} $internet_{n.sg}$}	TDC (1)
9.88	**hotline til problemer på internet** 'hotline to problems on (the) internet' {N2: $hotline_{n.sg}$ til_{prep} $problemer_{n.pl}$ på_{prep} $internet_{n.sg}$}	TDC (1)
9.89	**Kontakt TDC driftsproblem** 'Contact TDC operational problem' {N2: $Kontakt_{v.imp}$ $TDC_{n.prop}$} {T1: $driftsproblem_{n.sg}$}	TDC (1)
9.90	**løs problemer** 'solve problems' {T1/N2: $\text{løs}_{v.imp}$ $problemer_{n.pl}$}	TDC (1)

Summing up, what a query refers to depends on a variety of factors such as the task to be solved and the linguistic resources available to describe the various stages of the intended process.

As can be seen from the short collection of scenarios above, the choice of target primarily indicates whether the user needs much advice and suggestions or will prefer to simply be given what he or she is asking for. Thus, sophisticated methods for circumventing what is asked for and attempting to provide something that is helpful in spite of users asking for something else are probably best avoided when Stage 2 is targeted, but might potentially be very helpful with Stage 1 queries where users are typically in need of guidance and inspiration. The L^{WID} log contains a large amount of Stage 1 searches such as queries that consist of nothing but the user's nationality. Such users are in great need of inspiration, as they do not seem to know what jobs are available and suitable for them. Similarly, academics who use the site tend to use queries consisting of the name of the field they are working in, or the title they have achieved, and they apparently tend not to know what jobs to search for if that does not provide any useful results. The users responsible for these Stage 1 searches might well profit from a more inspirational approach which would guide the user through the possibilities and the job types available to them, whereas those searching directly for particular jobs do not need such functionalities to the same extent and would be likely to often find them more irritating than helpful.

9.3.5 Referring to actions

Some queries refer to an action that the user wants carried out, either by the system or by users themselves after having received information on how to do so. In the latter case, the searches are topical. Topical queries always belong to Stage 1, because what such queries refer to is a topic label, and this is of course known and present in the user's mind before the search is carried out. What is lacking is further information on the topic in question. Consequently, the topical type of action-referring queries also belongs in Stage 1.

9.91	**jeg er utilfreds** problem (Stage 1) {jeg$_{pron}$ er$_{v.pres}$ utilfreds$_{n.sg}$}	'I am dissatisfied'	TDC (1)
9.92	**utilfreds kunde** problem (Stage 1) {utilfreds kunde$_{n.sg}$}	'dissatisfied customer'	TDC (2)
9.93	**opsigelse privatkunde** solving action {(opsigelse$_{v.nom}$) (privatkunde$_{n.sg}$)} / {opsigelse$_{v.nom}$} {privatkunde$_{n.sg}$}	'termination (of subscription) private customer'	TDC (1)

On the other hand, some actions are ones that the system is expected to carry out for the user, and then it is so closely linked to the outcome as to be almost one. In such instances, referring to the action does not differ significantly from referring to the outcome, in the sense that both are examples of Stage 2 reference.

9.94	**print regning** solving action {N2: print$_{v.imp}$ regning$_{n.sg}$}	'print bill'	TDC (2)

Note that what is considered to be *the system* in this context is the combination of the search engine and everything else that the user subsequently interacts with in the process; thus, the action does not have to be one that the search engine itself can perform. This is important and seems to be true in general; people's behaviour does not appear to reflect any distinction between interacting with the search system and the larger system to which it provides access.

Some users simply refer to the actual resource that they are hoping to find by *anticipating* potential text in an imagined (or previously seen) resource. One strategy that seems to be especially common is to anticipate headlines of the type that uses an imperative to indicate that the ensuing text is going to tell the user how to do something particular. In some cases one may suspect that imperatives in queries are merely used because of their minimal length, but there is one kind of queries that proves the existence of **anticipated headlines** beyond doubt; some queries can only be interpreted as belonging to that type because they contain second person pronouns referring to the speaker.

9.95	**betal din regning** 'pay your bill' {T2: $\text{betal}_{\text{v.imp}}$ din $\text{regning}_{\text{n.sg(2p)}}$}	TDC (300)

It would be meaningless for users to refer to themselves in the second person, and so the use of second person forms must be a case of document authors referring to potential readers in this way in an anticipated stretch of text. L^{TDC} supplies 2374 examples of users reproducing a second person pronoun (*dig*, *din*, *dit*)[208] in their queries after an initial verb, spread over 34 distinct requests. In all but one of these requests the verb is in the imperative (in the remaining one it is in the infinitive). One request (representing 29 searches) even strengthens the anticipated character of the query by including a proximal pronoun *her* 'here'.

9.96	**betal din regning** **betal din regning her** 'pay your bill here' {T2: $\text{betal}_{\text{v.imp}}$ din $\text{regning}_{\text{n.sg(2p)}}$ $\text{her}_{\text{adv.prox}}$}	TDC (296) TDC (29)

Thus, there clearly are cases of anticipated text in the log material. However, most queries are not of this kind. In fact, there are even cases that cannot possibly be interpreted as anticipated text since they contain first person pronouns.

[208] Plural second person pronouns (*jer*, *jeres*) do not occur in the dataset; neither does the plural form *dine* of the singular possessive pronoun. The latter observation must be coincidental, but the former makes sense; sites rarely speak to their customers in the plural (this is of course not apparent in English).

9.97	**betale min regning** 'pay my bill' {T1: $\text{betale}_{\text{v.inf}}$ min $\text{regning}_{\text{n.sg(1p)}}$}	TDC (29)
9.98	**betal min regning online** 'pay my bill online' {T1/N2: $\text{betal}_{\text{v.imp}}$ min $\text{regning}_{\text{n.sg(1p)}}$ $\text{online}_{\text{adv}}$}	TDC (144)

L^{TDC} comprises 1333 searches where the queries contain first person pronouns (*mig*, *min*, *mit*, *mine*; no plural first person pronouns), corresponding to 22 distinct requests[209]. While ex. 9.97 is clearly a topical search for information on how to pay one's bill, ex. 9.98 is ambiguous between this interpretation and the more likely case of referring to the actual carrying out of the action as part of the interaction with the system. The use of the imperative seems to weigh towards this latter interpretation.

There is another use of imperatives which may well strike the English-minded reader as unexpected. The data show the existence of a special type of query which we shall call the **reproduced tool name** type. It involves reproducing the name of a tool or the text on a button or link, etc., for the purpose of learning how to use the tool in question. Text on buttons and links is in fact often expressed in the imperative, inviting the user to perform the act by clicking the button or link in question. In English, this may not be felt to be the case as there is no morphological difference from the infinitive, but Danish makes a clear distinction here[210]. Queries such as *send mail* belong in this category to the extent that the users intended to refer to the text on a button or link in order to have its uses explained.

[209] It should be noted, however, that the commonest of these, *ring mig op* 'phone me', at 464 searches, is an actual request in the normal sense of the word, asking as it does the reader (i.e. the company) to perform the action. Hence, it probably does not belong in this category. The same can be said of *kontakt mig* 'contact me', at 35 searches.

[210] Thence the problem of translating the command text *like* on social sites to other languages in which the imperative is not identical to the infinitive, and especially in the case of Danish where imperatives are often phonologically awkward; there is simply no imperative form of LIKE available in the Danish language, as such a command is not called for in normal language use. Hence, a special and rather awkward imperative form *synes godt om* which is indeed identical to the infinitive has had to be invented for use in such circumstances.

9.99	**send mail** 'send e-mail' {T1/N2: $send_{v.imp}$ $mail_{n.sg}$}[211]	TDC (118)

Reproduced tool names would seem to belong with topical queries; they represent something on which the user needs information. They are in a sense the opposite of anticipated headlines.

Ex. 9.99 may alternatively have been intended as a reference to the action of sending a mail. The importance is not to be able to classify each query unambiguously but to be aware of the different possibilities of interpretation. This type of queries can then simply be said to belong (unambiguously!) to an inherently ambiguous type. Again, tool names may actually contain second person pronouns (cf. *registrer dig* 'register (yourself)').

There is an interesting difference between tools that let users perform an action and those that allow them to initiate an action performed by the system; when a user presses a *cancel* button, it may be felt more as a request from the user to the system to perform the action in question, whereas if the button says *register* it is more of a request from the system to the user. However, the difference is purely deictic and therefore not reflected in the expressions, except when second or first person pronouns are included. This type is not likely to occur very frequently in search because in such cases the user would rather need to search for information on how to perform the action instead of trying to bring it about immediately. It would presumably not be possible, for instance, to use the imperative **registrer mig* 'register me' in reference to a desired action, since the verb is reflexive and hence must be in the second person in an imperative. Rather, the infinitive *registrere mig/sig* 'register (myself/oneself)' would be preferred in reference to a topic on which information was desired.

In conclusion, then, actions may be referred to in various ways, either as a topic on which information is sought, as anticipated headlines, or as an actual action carried out as the result of interacting with the system. The difference is clearly marked in some cases through the use of indexical and deictic terms, but in most instances only the semantics of the verb can tell us what was intended.

9.3.6 Interpreting a task

We now choose a case in which the Motivation is a sense of discontent. Some users will air this directly by referring to the Motivation.

[211] Note that this query is likely to be in Danish and therefore cannot be interpreted as involving an infinitive, which would be *sende*.

9.100	**utilfreds kunde** 'dissatisfied customer' {C1/S1: utilfreds kunde$_{n.sg}$}	TDC (2)
9.101	**jeg er utilfreds** 'I am dissatisfied' {C1/S1: jeg er utilfreds}	TDC (1)

These may be interpreted either as cases of Motivation or Source reference, depending on whether the discontent itself or the need to contact the company in order to complain is seen as the central problem (fig. 9.11). The choice is one that users are likely to have to make in the course of the search process as part of the reduction of uncertainty, and the viewpoint adopted probably has consequences for user behaviour, because the various targets change and the relationship between expression and intention alters with them.

	Problem interpretation 1	**Problem interpretation 2**
C1 Motivation	Discontent	Need for a satisfactory deal
S1 Source (problem)	Lack of contact information	Discontent
N1 Trigger	E.g. Need of a phone no. to TDC	E.g. How to contact TDC?
T1 Subject	E.g. Phone nos to TDC	E.g. Contacting TDC / Complaining
T2 Object	E.g. Phone no. to TDC	E.g. *Contact TDC* page
N2 Aim	E.g. Retrieve the phone no. to TDC	E.g. Find *Contact TDC* page / Call TDC
S2 Goal (resolution)	Having found contact information	Talk to TDC
C2 Objective	Ability to contact TDC	Relieved feelings / Better deal

Fig 9.11. Most tasks can be interpreted in multiple ways depending on what is taken to be the central problem. This reflects the layered character of tasks as reported by several authors (Xie, 2007).

The log contains no cases of people describing their lack of a telephone number to TDC, but this is solely due to the type of problem, which is unusually solution-focused. It is simply so much more natural to describe what is needed than what the problem is in this instance, as explained previously. This does not mean that the interpretation that this is the central problem is illegitimate; it is simply the case that people who choose this viewpoint tend to target the Goal rather than the Source. Apart from targeting the Motivation or Source, another possibility is to frame the Trigger. As what is needed is a piece of information, this could be done by posing a question.

9.102	**telefonnummer til tdc?** ‘phone number for tdc’ {N1: $telefonnummer_{n.sg}$ til_{prep} $tdc_{n.prop}$}	TDC (2)
9.103	**hvad er telefonnummert til tdc** ‘what is [the phone number] for tdc’ {N1: hvad er telefonnummert til tdc}	TDC (1)
9.104	**hvordan klager jeg** ‘how do I complain’ {N1: hvordan klager jeg}	TDC (12)

Next, it is possible to present it as a topic on which information is needed (the Subject).

9.105	**klage** ‘complain(t)’ {T1: $klage_{v.inf/n.sg}$}	TDC (3801)
9.106	**klage over regning** ‘complain(t) of bill’ {T1: $klage_{v.inf/n.sg}$ $over_{prep}$ $regning_{n.sg}$}	TDC (47)

Furthermore, direct reference can be made to the Object needed, such as a phone number or a guide to complaint-making.

9.107	**telefonnummer til tdc** ‘phone number to tdc’ {T2: $telefonnummer_{n.sg}$ til_{prep} $tdc_{n.prop}$}	TDC (170)

9.108	**klagevejledning** 'complaints guidance' {T2: klagevejledning$_{n.sg}$}	TDC (24)

Alternatively, text may be anticipated from the ideal page sought for.

9.109	**sådan sender du en klage** 'this is how you send a complaint' {T2: sådan sender du en klage}	TDC (7)

The Aim may be targeted by using an imperative.

9.110	**ring til tdc** 'call tdc' {N2: ring$_{v.imp}$ til$_{prep}$ tdc$_{n.prop}$}	TDC (147)
9.111	**send klage** 'send complaint' {N2: send$_{v.imp}$ klage$_{n.sg}$}	TDC (21)

The Goal in this scenario is a situation in which the phone number or the mail to which complaints should be sent has been procured. As with the Source, this is not a target that is normally used with this particular problem, though it is common enough with other types, the reason being that the Aim focus is the Object, and thus the Goal is simply an echo of the need associated with the Object, which is hence more likely to be targeted.

Finally, the Objective is to be able to call TDC or send a complaint, thereby giving vent to frustration and perhaps even ensuring a better deal with the company.

9.112	**bedre mobilabonnement** 'better mobile subscription' {C2/S2: bedre mobilabonnement$_{n.sg}$}	TDC (1)

With a differently framed problem this might also have been a Goal in itself, but in our scenario it was merely the Objective. This ambiguous interpretation of problems is typical and makes query analysis highly demanding and requires the analyst to be prepared to take great care not to force conclusions about how a given query is to be interpreted. Rather, inherently ambiguous classes which encompass the ambiguity in question should be erected in order to cover such instances. The framework presented above is the much-needed tool that allows one to begin this process by elucidating what the dangers are and making clear what targets there are to choose from in the first place.

9.4 Immediate and delayed action – Testing hypotheses

It is hypothesized here that, *mutatis mutandis*, when users search for ways of performing an action, they tend to use the imperative if they expect to be able to perform it directly as an **immediate action**, but prefer infinitives or nominalizations if they just want information on how to perform the action as a **delayed action**. Thus, actions such as making a call, which necessitate delayed action tend to be expressed in other ways than the imperative, while requests regarding actions that can be performed online such as registering for a service are more likely to be expressed in that form. It is not quite as simple as that, however; there are many countertendencies that blur the picture such as users *anticipating* phrasings in the imagined texts that they hope to find, overrepresentation of fixed or very common expressions, and overrepresentation of orthographically shorter or morphologically simpler forms, to mention some of the most obvious countertendencies. The result of this complex picture of interacting tendencies is that in absolute numbers it is in fact more common for the request *ring(e) til udlandet* ‘make a call to another country’ to be expressed in the imperative than in the infinitive, but we shall see that the overall tendency is towards compliance, when the effects of countertendencies are kept to a minimum. Thus, compared to the calling example, a request such as *opret(te) mailadresse* ‘create e-mail address’ has a much *stronger* preference for the imperative due to its greater likelihood of being perceived as a potential immediate action.

In the following examples, the imperative variant is likely to indicate that the user believes (consciously or subconsciously) that it is possible to register onsite, while the infinitive signals a desire to receive information on how to register.

9.113	**tilmeld betalingsservice** ‘register (for) payment service’ {N2: tilmeld$_{v.pret}$ betalingsservice$_{n.sg}$}	TDC (13)

9.114	**tilmelde betalingsservice** 'register(ing) (for) payment service' {T1: tilmelde$_{v.inf}$ betalingsservice$_{n.sg}$}	TDC (1)

It is difficult to prove that this is so, except by introspection and logical inference, but the empirical study presented below does at least indicate very strongly that there is a meaning difference between imperatives on the one hand and infinitives and nominalizations on the other when used in queries. Exactly what the difference is can only be guessed at unless a full-scale behavioural study is carried out, but we shall suggest tentatively that it is one of *immediate* vs. *delayed Aim*[212].

9.4.1 Tendencies and countertendencies in search behaviour

The following are some of the most important superficial tendencies that may be presumed to influence the choice of form when query terms are expressed.

- Markedness of form
- Shortness of expression
- Fixed expressions
- Anticipated text

As our goal here is to identify the basic principles that determine the coding of intentions in queries, these countertendencies will have to be filtered away as far as possible. We shall thus be studying the intentional coding by attempting to keep the effect of the countertendencies to a minimum. Each countertendency is dealt with in turn below.

Tendency 1: Markedness of form

Basic forms of lexemes might be assumed to occur more commonly in queries since they represent the concept by way of naming it. This may or may not be true, but in any case such a tendency is hard to distinguish from the tendency towards shortness of expression. I do not actually believe there to be any significant effect of the markedness of form tendency, if one thinks of it as a propensity to use the lemma form of a word. On the other hand, there is

[212] As mentioned above, a transaction that takes place in immediate succession of a search is typically treated as part of the search for the purposes of Aim calculation, since the two tools are considered aspects of one. Consequently, users may aim to acquire an object or solve a problem despite being fully aware that this cannot be attained through simple use of the search engine. Immediate problem solving thus includes any tasks that can be performed on site in succession of the search.

indeed a strong tendency not to use very marked forms, such as present and past forms of verbs, or definite and genitive forms of nouns. Such forms are likely to be used principally as an effect of tendency 4, in anticipation of resource text. Verbal forms are typically cited in infinitive, imperative, gerund or nominalized form[213], and nouns in either singular or plural, depending on the interaction between intention coding (cf. tendency 5) and the effect of countertendencies.

● Tendency 2: Shortness of expression

For reasons of economy, shorter forms of words are likely to be preferred whenever possible, to the extent that this does not interfere unnecessarily with intention coding. Shortness of expression can be achieved in several ways such as by a) avoiding affixation, b) abbreviating words, c) clipping words, and d) leaving out part of a compound or phrase. The following is an example of the latter type.

9.115	**strawberry**	WID (64)

As the user can hardly be imagining an interaction with only one strawberry, this is likely to be a reduced version of a compound such as the following with which it co-occurs in the log.

9.116	**strawberry picking** {S2: strawberry picking$_{v.nom}$}	WID (10)
9.117	**strawberry picker** {S2: strawberry picker$_{n.sg}$}	WID (6)

The reduced version is by far the more common, and it is also more frequent than the plural form.

9.118	**strawberries** {S2: strawberries$_{n.pl}$}	WID (40)

[213] English does not distinguish between infinitives and imperatives, and Danish does not have gerunds but uses infinitives or nominalizations instead.

Of course, this might be thought of as an example of a tendency for using the least marked form, but as the WID domain ensures that these users are not looking for information on strawberries, it is more likely to be a case of shortening a compound.

Clipping words is extremely rare in the analysed logs, and when it does occur, such spellings are more likely to represent typos. In fact, there is a strong tendency to use full and rather complex phrases; most likely this is because people have a natural and deeply rooted aversion against incomplete words and phrases which carries over into query formation.

Abbreviations do occur in the logs, especially in established cases such as *tlf* for *telefon*. This is in fact surprisingly frequent given that a LAT cannot *a priori* be expected to be able to interpret such expressions. In L^{TDC}, 6196 searches use the string *tlf*, even as part of compounds which are written solid.

9.119	**tlf** **tlf.** {N2: $tlf_{n.sg}$}	‘tel.’	TDC (612) TDC (29)
9.120	**tlf nr** **tlf.nr** **tlf.nr.** **tlf. nr.** **tlf. nr** **tlf nr.** **tlfnr** **tlfnr.** **tlf nummer** {N2: tlf $nr_{n.sg}$}	‘tel.no.’	TDC (499) TDC (139) TDC (66) TDC (52) TDC (42) TDC (41) TDC (88) TDC (15) TDC (132)

9.121	**mobiltlf** **mobiltlf.** **mobil tlf** {N2: $\mathrm{tlf}_{n.sg}$}	'mobile tel.'	TDC (76) TDC (14) TDC (86)
9.122	**spærring af tlf** {T1: $\mathrm{spærring}_{v.nom}$ af_{prep} $\mathrm{tlf}_{n.sg}$}	'blocking of tel.'	TDC (24)
9.123	**mailadr eller tlf nr til tdc** {N2: $\mathrm{mailadr}_{n.sg}$ eller_{conj} tlf $\mathrm{nr}_{n.sg}$ til_{prep} $\mathrm{tdc}_{n.prop}$}	'e-mail addr. or tel. no. to tdc'	TDC (5)
9.124	**tlf.regning** {N2: $\mathrm{tlf.regning}_{n.sg}$}	'tel. bill'	TDC (3)

● **Tendency 3: Fixed expressions**

Some words are either fixed in one form, or are simply much more frequent in that form. For instance, the form *indstillinger* 'configurations' (1751 occurrences) is far more common in L^{TDC} than *indstille* 'configure' (59 occurrences), presumably because people tend to think of configurations not as something you do (i.e. configuring) but rather as a set of parameters. Even the singular form of the nominalization *indstilling* 'configuration' is much rarer, at 454 occurrences.

● **Tendency 4: Anticipated text**

It is hardly surprising that people should tend to anticipate text in the potential resources that they are looking for. After all, with primitive search systems that was the only viable approach. This tendency probably accounts for the presents and pasts, the definites and the genitives, etc. It also accounts for the choice between less marked forms in many cases, such as imperatives thought (or hoped) by users to occur in headlines and on buttons and links, such as *become a member*, *contact us*, and *print this page*. The text that people anticipate may reveal something of their intentions, but on the other hand there is no need to understand the intention, since the documents are likely to be found anyway, as the text (if indeed it occurs) is stated *verbatim* in the query (though precision may of course be a problem in some cases). It should always be kept in mind that such literal search remains highly useful in some cases, and that users should be allowed to carry it out whenever needed, without interference from sophisticated interpretational algorithms.

The tendency for anticipation is likely to contribute to the high frequency of imperatives in the mentioned query *ring til udlandet*, especially as the effect is likely to be active without the user being fully aware of it, which means that not only the most probable headers are anticipated.

All of the above tendencies are rather superficial from the point of view of intent recognition, i.e. they do not really represent any differences in intention. In contrast, the following tendencies may be considered "rules" or principles of intention coding, or in other words they represent real referential or intentional differences. Tendencies do not have an absolute strength hierarchy but appear to be unpredictable and fluctuating effects (due to the cybernetics involved), so that the outcome is never entirely predictable. The best we can do is interpret the resultant query afterwards. Hopefully, we can understand what made the user express a certain need in that particular way by comparing the potential effects of the various tendencies which might have been involved. The rules that are hypothesized to occur relate to *extensional* and *stadial* differences.

● **Tendency 5: Extensional differences**

In some cases, there is a real referential difference between forms that mirrors the referent in the world referred to, as with some instances of plural nouns as well as aspectual differences in verbs. This would account for the relatively frequent occurrence of the plural form *strawberries* in L^{WID} (40 as against 64 singulars); because people expect to interact with multiple strawberries in their would-be job, it is natural to use the plural form, unless other tendencies are stronger in that moment.

● **Tendency 6: Stadial differences**

Stadial choices have slightly different effects with nouns and verbs (including nominalizations of verbal roots), but the basic principles are the same. Unlike extensional differences, stadial ones are mostly not morphologically marked.

9.125	**selvbetjening ny bruger** 'self-service new user'	TDC (1)

It is impossible to tell whether this query refers to the user as a currently new customer or to the state that the user wants to enter, i.e. a future new customer; it could be paraphrased as either *I AM A NEW CUSTOMER AND I WANT TO START USING SELF-SERVICE*, or *I WANT TO USE SELF-SERVICE IN ORDER TO BECOME A NEW CUSTOMER*.

With verbs, users may refer to either something that is currently happening, an action that they want to perform, or a kind of activity that they need information on, often in order that they may carry out the action in the near future.

9.126	**iphone låst** 'iphone locked' {S1: $\text{iphone}_{n.sg}$ $\text{låst}_{n.sg}$}	TDC (5)
9.127	**lås iphone op** 'unlock iphone' {S2: $\text{lås\_op}_{v.imp}$ $\text{iphone}_{n.sg}$}	TDC (32)
9.128	**oplås iphone** 'unlock iphone' {S2: $\text{oplås}_{v.imp}$ $\text{iphone}_{n.sg}$}	TDC (19)
9.129	**oplåsning af iphone** 'unlocking of an iphone' {T1: $\text{oplåsning}_{v.nom}$ af_{prep} $\text{iphone}_{n.sg}$}	TDC (18)
9.130	**låse iphone 4 op** 'unlock(ing) iphone 4' {T1: $\text{låse\_op}_{v.inf}$ $\text{iphone 4}_{n.sg}$}	TDC (4)

Ex. 9.126 refers to Stage 1, the locked phone, ex. 9.127-9.128 to the act of unlocking it as something to be carried out now[214], and the users responsible for ex. 9.129-9.130 seem to be looking for information on how to unlock their phones. A fuller study of this case is included in Appendix E2.

[214] This instance can alternatively be explained as the effect of a countertendency, e.g. anticipation of headlines.

9.4.2 Action classes

In order to test the composite hypothesis of tendencies and countertendencies, L^{TDC} was analysed. The following types of intended responses were predicted to occur in the dataset, correlating with different preferences for certain morphological forms.

● **Action Class 1: Immediate (user) action**

A user expecting to be able to perform the action directly on site or in immediate succession of the search tends to use an imperative. A typical example is the imperative *find* ('find'), where the action is carried out by the system, but the user is the agent using the system as an instrument. Here, imperatives are used almost exclusively.

● **Action Class 2: System action (Passive user)**

In some cases, users do not look for ways of performing an action (neither on their own nor through the system), but rather intend the system to perform the action for them. This is likely to be expressed through an imperative. One example would be the imperative *oplys* (inform), which is a request for the directory service, called in Danish *oplysningen* ('the information'). Leaving aside the fact that the nominalization is strongly predominant due to its lexicalized status, the only other form encountered is the imperative (215 of 82,540 queries).

● **Action Class 3: Delayed (user) action**

A user who does not expect to be able to perform the action directly on site or in immediate succession of the search, but rather is looking for information on how to perform the action afterwards, tends to use an infinitive or a nominalization. An example would be complaining about services, which is typically done separately by dropping a line or sending a letter or an e-mail. Here, less than 2% of users employed an imperative, while the remaining 98% of queries were all ambiguous between the infinitive and the noun *klage* ('complain' vs. 'complaint').

● **Action Class 4: Named action (Reproduced tool description)**

If the user is looking for help with application tools and functionalities, the reverse of anticipation may occur, i.e. the user may reproduce the name of the tool, which in Danish (and English) will often include an imperative[215]. As these are cases of delayed action, this method can be said to counteract the tendency to conform to action class 3. Nevertheless it seems reasonable to designate it as an action class in its own right. Note that tool descriptions are very often restricted to an imperative with no context added so that this tendency is in

[215] In Romance and Indian languages it is common for infinitives to be used in such cases.

most cases unlikely to be at play when a context is provided in the query, exceptions being examples such as “write new mail”. Sometimes, tool descriptions may indeed be anticipated, if users search in the hope of finding a button or functionality with the name in question. This type of search will be grouped with anticipated headlines under countertendency 1 below.

9.4.3 Countertendencies

The following countertendencies were predicted to affect the distribution of forms in the dataset, thereby blurring the picture as regards the correlation between action classes and morphological preferences.

- **Countertendency 1: Anticipated headline**

Users who are looking for instructions on how to perform actions may anticipate the use of imperatives in headlines above such instructions, such as *pay here* or *cancel your subscription*. This will tend to counteract the tendency to conform to action class 3. A typical example of this is queries representing a desire to be told how to write to the company. Here, almost 94% used the imperative *skriv* (‘write’) in spite of the fact that it represents a delayed action; only 6% used the infinitive *skrive*. This is likely to be caused by an anticipated headline along the lines of *skriv til os* (‘write to us’).

- **Countertendency 2: Anticipated text**

Users may anticipate actual expressions in text on site, leading to the use of tensed forms as well as other verb forms and nominalizations, and this may counteract other tendencies. The method is probably more likely to be used with delayed action than with immediate action, since the user is explicitly looking for a text rather than the possibility of performing an action. Consequently, tensed forms are here grouped with infinitives and nominalizations rather than with imperatives. Sometimes, a text may contain imperatives, and users may of course anticipate these as well. However, it was decided that anticipated imperatives should by definition be grouped with the headlines in countertendency 1.

- **Countertendency 3: Fixed nominalizations**

Some expressions are so overwhelmingly used in nominalized form that verbal forms are significantly less likely to be used. This counteracts any other tendencies to the contrary. In L^{TDC}, this is the case with *oplysningen* (‘directory service’). Other examples are *kryptering* (‘encryption’) and *indstillinger* (‘configuration’) which are both used in 97-98% of the queries referring to those verbal roots.

● **Countertendency 4: Shortness of expression**

Users are likely to prefer the shortest possible expressions whenever possible. In Danish, as in many other languages, this is the imperative. This predilection may thus counteract any other tendencies to the contrary. This is in fact likely to be the main reason for the presence of the imperative *oplys* ('inform') in the corpus, as it seems somewhat farfetched to use an imperative for other reasons in this particular case. It is difficult to find clear instances as there are always alternative explanations for the use of an imperative. However, at least in cases where there is an overwhelming tendency to use an infinitive, such as *reparere* ('repair'), it can be reasonably supposed that the few imperatives that do occur (in this case less than 5%) are instances of preferring the shortest expression possible.

● **Countertendency 5: Phonological awkwardness**

In Danish, certain imperatives such as *ændr* are phonologically awkward and tend not to be used. In these cases, therefore, users are likely to prefer the infinitive form (*ændre*), though whether they consider this to actually be a representation of an imperative or an infinitive is unclear. This phenomenon will counteract tendencies to use an imperative for those particular verbs whose stems happen to end in a plosive followed by a sonorant. Other examples from the corpus are *åbne* ('open') and *genåbne* ('re-open'), neither of which occurs in the (orthographically correct) imperative at all.

● **Countertendency 6: Systematic spelling mistakes**

In Danish, there is a tendency to confuse the infinitive and the present of verbs ending in *-ere*. This means that present forms of such verbs are probably more likely to represent infinitives. The opposite may of course also occur. The corpus contains the present forms *abonnerer*, *aktiverer*, *deaktiverer*, *fungerer*, *installerer*, *korrigerer*, *signalerer*, and *vurderer*, all of which were most likely intended as infinitives.

These tendencies may be considered in light of a hypothesized basic distinction between immediate and delayed actions blurred by the other, counteracting tendencies. In particular, class 3 (delayed action) will be blurred by countertendency 1 and 4, while class 1 and class 2 may be counteracted by countertendency 3 and sometimes countertendency 2.

9.4.4 Data analysis

All queries in L^{TDC} with an initial verb or nominalization based on a verb were extracted and tagged with their morphological form using Ankiro Thesauri, part of the knowledge base on which Ankiro's semantic search engines are built. There were 293,960 searches of this kind, spread over 1631 different query strings. The queries were sorted by the verbal root they contained. Some forms were ambiguous between imperatives and nouns, e.g. *test* and *download*. In order to get the most clear-cut results possible, all such queries were excluded from the statistical analysis. Furthermore, only verb roots that occurred at least 50 times were counted. These trimmings did not materially affect the outcome of the analysis, but they made the tendencies clearer and the results more reliable.

Some verbs occurred in various contexts and could be split into several senses of the orthographic verb. Regardless of whether different meanings are considered variants of one lemma or different words altogether, it made sense to split them into separate analytical entities because user behaviour is likely to differ between divergent uses of the same verb as intentions vary accordingly. The clarity of results improved significantly with this disambiguation. The final analysis contained 252,495 queries spread over 102 different verbs senses (fig. 9.12). Appendix E3 contains the full data as regards the actual verbs.

1	oprette (service) – create (service)	35	godkende – approve	69	overføre – transfer
2	søge (nummer) – search (number)	36	aflytte – listen (to voicemail)	70	fortryde – cancel (action)
3	skrive (mail) – write (e-mail)	37	betale – pay	71	flytte – move
4	hente (indhold) – open (contents)	38	slette – delete	72	opsætte – configure
5	fjerne – remove	39	byte – exchange	73	optanke – replenish
6	tilknytte – assign	40	vælge – choose	74	fejlmelde – report (error)
7	blive – become	41	oprette (hjemmeside) – create (homepage)	75	indtaste – type
8	indsætte – insert	42	melde – report	76	synkronisere – synchronize
9	redigere – edit	43	leje – hire	77	reparere – repair
10	sammenligne – compare	44	bastille – order	78	viderestille – divert
11	videresende – forward	45	beskytte – protect	79	opsige – terminate
12	uddybe – elaborate	46	læse – read	80	opdatere – update
13	spole – rewind/fast-forward	47	gendanne – restore	81	omstille – divert
14	gøre – do	48	framelde – unsubscribe	82	klage – complain
15	henvende – contact	49	surfe – surf	83	overdrage – transfer
16	finde – find	50	afmelde – unsubscribe	84	oplyse – inform
17	skrive (klage) – write (complaint)	51	blokere – block	85	abonnere – subscribe
18	tømme – empty	52	opgradere – upgrade	86	tilslutte – connect
19	sende (mail) – send (e-mail)	53	installere – install	87	give – give
20	oprette (bruger) – create (user)	54	invitere – invite	88	fungere – work
21	tilmelde – register	55	lukke – close	89	vurdere – assess
22	spørge – ask	56	beholde – keep	90	opfylde – replenish
23	nulstille – reset	57	fremkalde – develop (photos)	91	indstille – configure
24	oprette (mail) – create (e-mail)	58	søge () – search ()	92	kryptere – encrypt
25	scanne – scan	59	registrere – register	93	genåbne – reopen
26	oversætte – translate	60	geninstallere – reinstall	94	besvare – answer
27	oprette (mailadr.) – create (e-mail address)	61	modtage – receive	95	åbne – open
28	ophæve – terminate	62	udskrive – print	96	sikkerhedskopiere – backup
29	følge – follow	63	afbestille – unsubscribe	97	begrænse – limit
30	etablere – establish	64	optage – record	98	misbruge – abuse
31	male – measure	65	påfylde – replenish	99	lokke – lure
32	sende (klage) – send (complaint)	66	aktivere – activate	100	ændre – change
33	udvide – expand	67	konfigurere – configure	101	klikke – click
34	deaktivere – deactivate	68	oprette () – create ()	102	hente () – open/get ()

Key to the columns in fig. 9.12. The full data are available in Appendix E3.

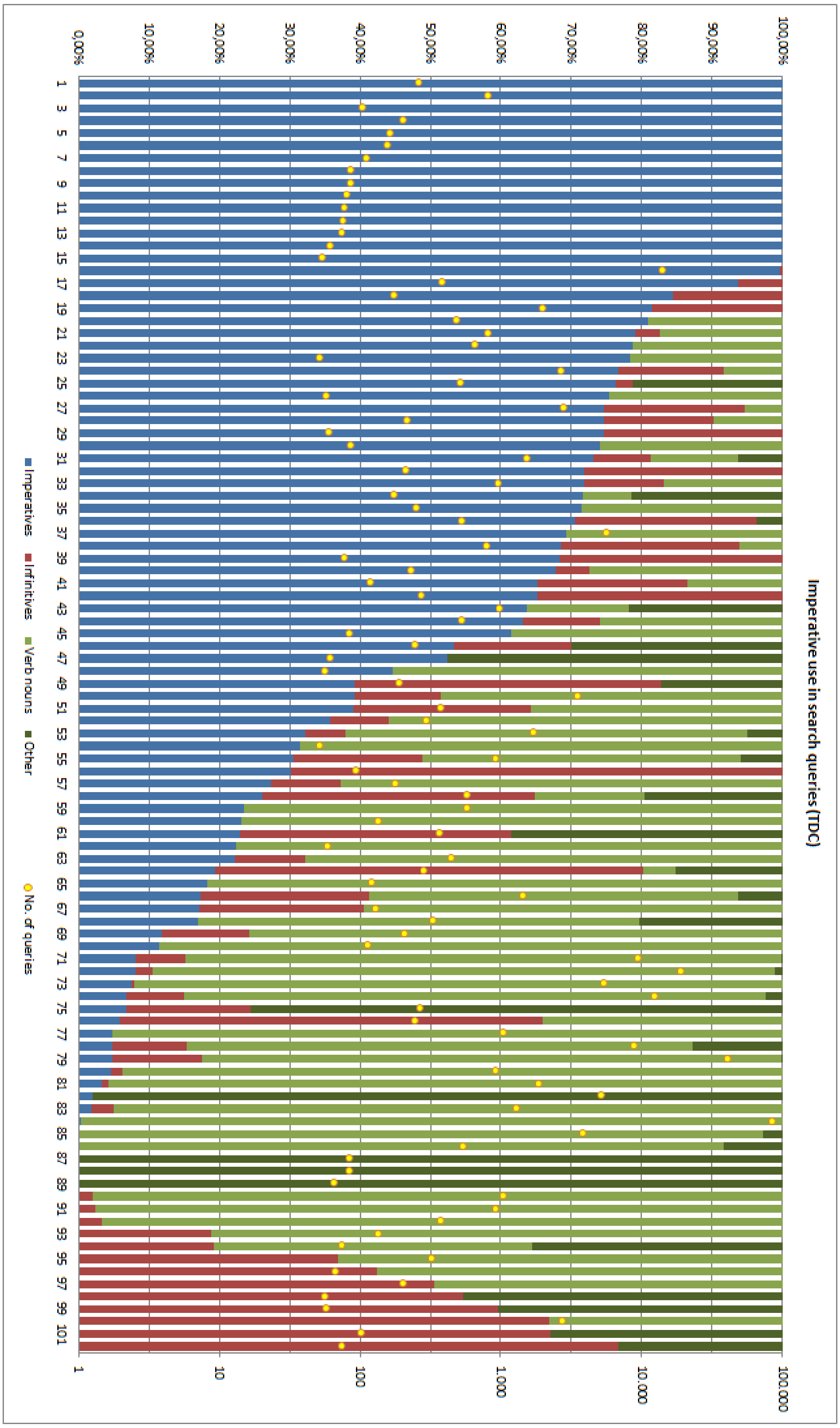


Fig. 9.12. The use of imperatives and other verb forms (including nominalizations) in L[TDC]. The x-axis contains different verbs and verb senses. The yellow dots indicate the (logarithmic) frequency of the verbs or verb senses as the case may be, and hence the strength of the results (the logarithmic scale was used due to the distribution of frequencies which becomes somewhat even with this representation).

9.4.5 Results and discussion

Although the nature of this study does not allow us to state anything with certainty, the results do suggest that there is indeed a tendency to conform to our hypothesis. As might be expected, the indications are strongest with verbs that are used very frequently.

It is clear that not all verbs behave in the same way. Some only occur as imperatives, some never do, some often, and some rarely. What is remarkable is that there are more or less just these four types of verb behaviour.

1. 100% imperatives
2. 60-80% imperatives
3. 1-40% imperatives
4. 0% imperatives

Apart from the gradual cline from 0% to 40%, there are almost no verbs that fall between the categories, and the question is why the behaviour of verb form use should display these clusters. The most likely interpretation is that it reflects a distinction between semantic verb types that are likely to be used in different circumstances, such as a distinction between *immediate action* searches and those that are more likely to be used in *delayed action* searches. The fact that the diagram does not just show a regular growth but is clearly divisible in separate groups suggests that this corresponds to distinct semantic types rather than displaying random variation. There are very few verbs that have an equal likelihood of occurring in both types of search; the zone between 40% and 60% is almost empty and is likely to reflect such a divide. The variation that *is* observable within the groups is a natural consequence of the fact that most verbs have more than one semantic interpretation and can hence be used in different contexts, in addition to the effects of the various countertendencies. For instance, the fact that imperatives are the shortest forms could well be the explanation of the cline between the two groups that do not favour imperatives. Thus, there are two main groups of 0-40% imperatives and 60-100% imperatives, respectively. The latter group mostly contains percentages lower than 80%, and the ones that are higher generally have low frequencies and thus may be less significant. On the other hand, many of those that never occur in the imperative have relatively high frequencies, which means that there is likely to be a real reason for the absence, such as the effect of one or more tendencies and countertendencies. This is the case with the verb *ændre*, for instance, which has a very awkward imperative *ændr*, and which consequently may be written in infinitive form, even when it is interpreted as an imperative.

Significance can be enhanced by only counting verbs that occur with a following object or prepositional phrase (fig. 9.13).

Contextualized verb		**Imperative use**
oprette (service)	'set up (service)'	100%
oprette (bruger)	'create (user)'	81%
oprette (mailadresse)	'create (e-mail address)	75%
oprette (hjemmeside)	'create (homepage)'	65%

Fig. 9.13. Comparison of the frequency of imperative use with different context-derived senses of the verb *oprette*.

This causes the tendencies to become even more sharply delineated (fig. 9.14). Now the middle break between the two types is unmistakeable – there are no verbs in the dataset that occur as imperatives between 49% and 64% of the time. This suggests a rather clear division between at least two behaviour classes. The groups that emerge are as follows.

1. 100% imperatives
2. 65-93% imperatives
3. 5-49% imperatives
4. 0% imperatives

The explanation that suggests itself is that adding context separates various meanings of the same orthographic verb further and hence clarifies the distribution. The difference between various senses of the same word may be very clear in some cases.

It is certainly also very clear that these are only tendencies; no absolute rules for the use of imperatives can be given. Indeed, if such rules of mapping between verbs and forms existed, it would not be half so interesting, because then it would merely be a case of lexical variation rather than a distinction between different intents.

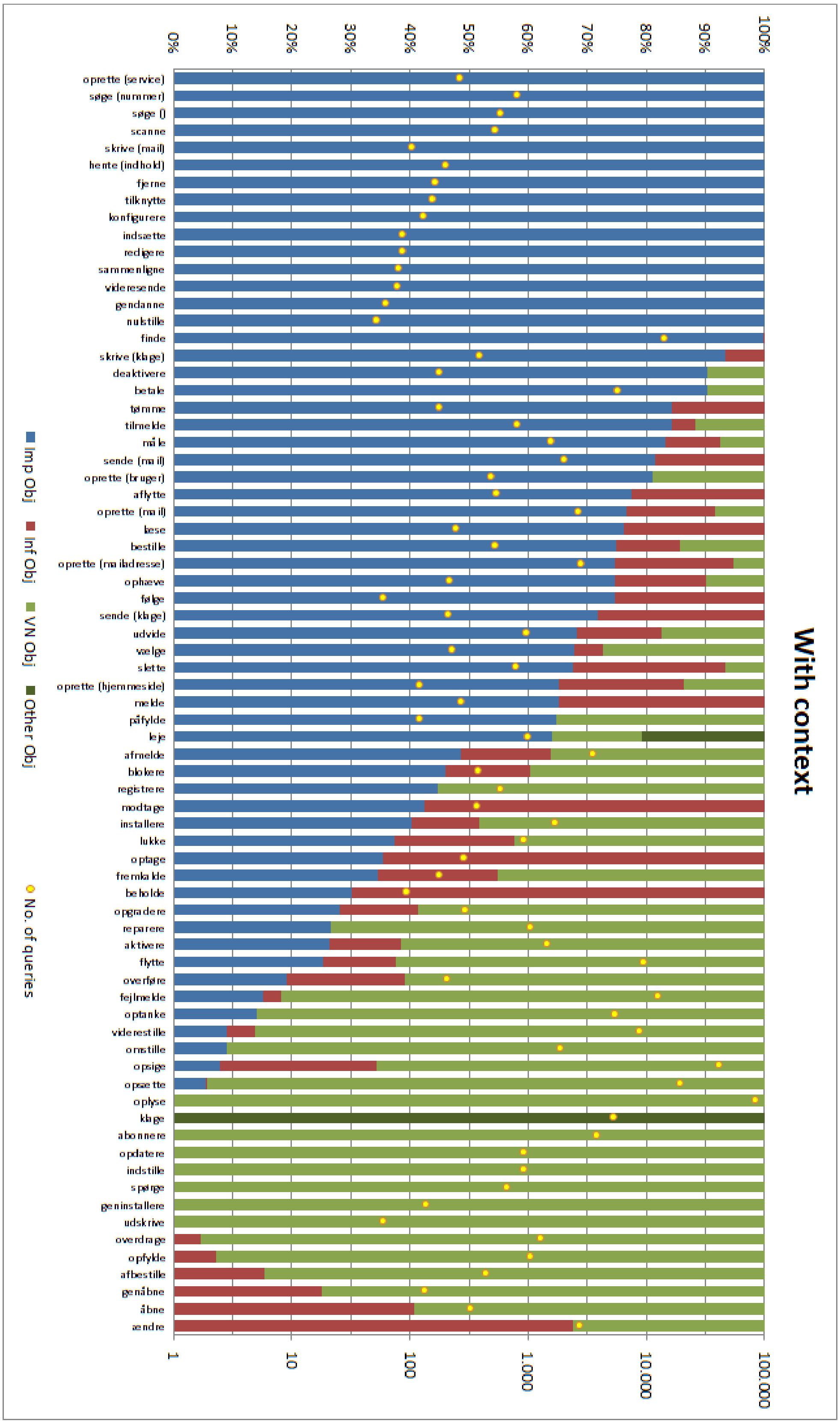


Fig. 9.14. The use of imperatives and other verb forms (including nominalizations) in L[TDC], when only verbs with a following noun phrase or prepositional phrase are included.

Summing up, the study indicates that not all verbs have an equal tendency to be used in the imperative, and this difference seems to some extent to be linked to the semantic sense of the verb. The frequencies cluster into at least two groups which may be interpreted as reflecting a differentiation in behaviour corresponding to the intention of wanting to have an action carried out in immediate succession of the search and desiring information on how to perform it later. However, this cannot be proved without further analysis involving questionnaires or similar references to the users' actual intentions, and even then it would require careful design of the study as users are probably not aware of their underlying reasons for using a certain form. Thus, we shall content ourselves with the tentative suggestions provided by the above study at present. Hopefully, the matter will be studied more thoroughly in future.

10 In search of a solution

> *The real-life issue in IR systems design and evaluation is not whether a proposed method or tool is able to improve recall / precision by an interesting percentage with statistical significance. The real issue is whether it helps the searcher better solving the seeking and retrieval tasks (faster, with less resources, with better result quality). This has to do with learning about the search task, formulation of the request, a variety of tactics.*
>
> (Ingwersen & Järvelin, 2005a)

It might be thought that intentions would be inherently unrecoverable from log files and could only be accessed by asking or in other ways extracting information from the users themselves on the fly. In support of this, Grimes *et al.* (2007) claim that "logs can only ever reveal correlations and not causality, however tempting inferring causality may be". This would certainly be true were the logs in question only records of click behaviour; you cannot determine *why* a user clicked a resource without resorting to guesswork. However, these authors are not reckoning with the fact that queries are linguistic messages and consequently do intrinsically convey intentions. This means that it should indeed be possible to some extent to identify intentions by attending to queries. What this requires is an appropriate framework of intention types that does not go beyond what is conveyed in queries (unlike the traditional types used for query classification), as well as strict adherence to what is actually expressed when performing the log analysis. Consequently, we need a notion of what the user intended that connects with the expression so that we can study only the kind of intention that is actually identifiable through query logs. The notion of targets comes in handy here by nuancing the way we look at intentions, and in addition we have found a need for a specialized notion of user *intent* that differs from the general *intention* in being explicated in a query. Query classification can only be truly useful to the extent that we abandon the urge to classify according to some typology of abstract needs or intentions that are not explicitly reflected in the query.

The important thing to remember in order to avoid faulty or arbitrary classification is to not shy away from defining inherently ambiguous classes, because an ambiguous class is much more informative than an arbitrary classification based on clear and unambiguous categories, and while it may not be possible to get clear and unambiguous answers from a search log, this does not mean that no information can be had. There is clearly a lot of variation in the expression of queries, evidence of the expressive choices made by users when formulating query strings, and these choices are likely to reflect intentions in certain complex ways. It is our job as IR researchers and designers to take advantage of this in future.

10.1 Target Analysis

Performing a linguistically informed **Target Analysis** requires the prior identification of a set of context-dependent categories which define the conceptual targets that users search for within the domain in question. Going beyond form should be limited to this description of the domain, so that the actual analysis can be done in an unbiased manner in accordance with the model, knowing that subjectivity in the analytical choices is limited to assumptions founded on the previously developed understanding of the domain and thus presumably rests on sound and rational principles. For an example of such a choice, consider job description queries again. Such queries may alternatively be construed as referring to a job that the user has held previously rather than desired future jobs, and in many cases even to an education aimed at such positions. These interpretations are certainly all possible. The ambiguity is regular in the sense that it affects all job descriptions equally (or at least all within a clearly delimited group, as in the case of ambiguous educations and jobs), and as such the set of queries referring to job descriptions may be treated together in a consistent manner. It is presumably reasonable for the analyst to make the decision that all job descriptions are analysed as referring to the desired job within the confines of a job search portal, unless there are explicit clues to the contrary, but the queries in this set are inherently ambiguous.

For another example, consider the frequent occurrence of personal names and phone numbers in L^{TDC}. Such queries by themselves refer simply to a special type (or types) of topic, viz. people. They accordingly belong to what is traditionally called informational searches. However, in the context of a telecommunications site they are in most cases almost certain to represent an intention of looking these people up in an online telephone directory, rather than being requests for topical information on them. As such, they are in fact more akin to transactional searches, as they extract dynamic information from a database, while at the same time being somewhat like the job description case in that the users are looking for a (dynamic) page representing the sought-for person. In other words, these searches are somewhere between informational, transactional and navigational queries.

Rather than trying to guess what the user intended, with Target Analysis intent classification becomes a matter of identifying possible (and likely) inferences from a simplified expression to a real-world situation in which it was issued as a query. For instance, we might try to analyse the following query.

10.1	**hotel**	WID (696)

What possible and likely relations can there be between the concept HOTEL and the situation in which a person using the Workindenmark site would issue such a query? There are several possible answers.

1. The user may want to work in a hotel, i.e. it refers to the **place** that defines the kind of work that the person is looking for.
2. The user may have worked in a hotel before and has chosen to mention this in the hope of finding a job that is suitable for someone with such a **background**.
3. The user may want to work with hotels as the **object or product** of the work, e.g. the construction of hotels.
4. The user may want to inform the system of his or her capabilities regarding hotels, i.e. as a **competency** relating to hotels in some way.

The point of a Target Analysis is not to estimate whether the user was most likely to be looking for particular resources but rather to determine whether users are most likely to *refer* to their background, the *place* they wish to work, the *object* that they want to work with or their *competencies*, or something else. Not all of these interpretations have an equal likelihood. In fact, it is probably reasonable once again to make the formal assumption that *hotel* searches are most likely to indicate the place in which the user is looking for a job, all other things being equal. This is a choice made by the analyst, but it is based on informed analysis of the context domain. In order to minimize bias, all searches which are ambiguous in the same way should be assigned to the same subtype so that they are kept as separate types, rather than, say, assigning all queries which may refer to characteristics of a desired job to one category and ignore the various types of ambiguities. Thus, in this case queries such as ex. 10.1 might be classified as instances of the category *workplace*. The analyst would then know that all workplaces are potentially ambiguous in the ways mentioned (though some may possibly be less likely in certain interpretations, for various reasons, and the analyst must decide on a granularity to apply). Sometimes, disambiguating information may have been added by the user, in which case the queries may indeed be assigned to another category.

10.2	**hotel job** {S2: hotel $job_{n.sg}$}	WID (19)

An interpretation as {S1: $hotel_{n.sg}$} {S2: $job_{n.sg}$} in which the user informs the system that he or she is qualified for work in hotels and is looking for a job is not impossible here, but the interpretation as a compound noun is probably more likely. Again, the ambiguity is unavoidable, but this should not deter the analyst from inferring any intentional information at all, as long as the analysis is kept consistent and rigorous and is based on a solid knowledge of the domain.

10.1.1 Modelling a domain – Preparing for a log analysis

None of the various definitions of the Broderian query types given in chapter 7 make any direct reference to the actual form of the query, although some of the authors do in fact seem to intend this. For instance, based on Saracevic's (1997c) stratified model Jansen *et al.* define *user intent* as "the affective, cognitive, or situational goal as expressed in an interaction with a Web search engine" (Jansen, Booth, & Spink, 2008, p. 1255). They explicitly include the expression in their definition because "the expression determines what type of resource the user desires in order to address their overall goal". This is hardly a suitable way of stating it, as one would expect intention to determine form rather than the other way round, but at least they speak of a connection between the two. In an attempt to live up to this, they base their criteria on superficial characteristics of queries. However, this does not assure faithfulness to users' intentions because the interpretation depends on context, which is not included in their analytical framework. To take one example, they classify "queries containing company/business/organization/people names" as representing navigational intentions, even though such queries have a range of uses, and informational intentions would certainly not be rare among them, especially with people's names. Thus, user intent arises in, and is inseparable from, context. The analysis of job description queries confirms that queries can only be analysed with reference to the context in which they were issued. Some going beyond form is indeed necessary in order to get anywhere, but only to the extent that the inferred can be recovered from context. In a way, such inferences *are* coded in the linguistic message, because a linguistic message *includes* the context in which it is uttered and can only be interpreted within that context. In other words, the query is not just a string but a message uttered in a context, and the string is only a cast of the *real* query. In order to understand the real query its context must be reconstructed. The reason that a search for a job posting can be accomplished by simply entering a job description is exactly because the context is known and therefore the user may – and normally does – exclude any superfluous explanations of what the intentions of such a search might be. If performed in a web search environment, a search of that kind would be likely to include more information, through the addition of words such as *job*, or *postings*, or something to that effect. In contrast, only a single user of the Workindenmark site thought this necessary.

10.3	**physician job** {S2: (physician) < (job$_{n.sg}$)}	WID (1)

The same expression may thus be used with several different intents depending on context. For instance, in L^{WID} the following query is likely to represent a wish to be picking strawberries in the desired future job position, rather than information on picking strawberries, whereas in a different context the interpretation would be entirely different.

10.4	**strawberry picking** {S2: strawberry$_{n.sg}$ picking$_{v.nom}$}	WID (10)

Queries are thus rarely sufficiently specific in themselves to reveal what kind of search they represent; it is only when taken in context that this can be guessed at with any certainty, and so before a log analysis can be made the domain needs to be modelled so that contextual clues can be drawn on. One site may accommodate many different scenarios, which need to be disentangled and modelled. The following are common types.

- **Problem solving**

  Problem solving typically involves a problematic Stage 1 and a possible solution Stage 2. In a prototypical problem, it is Stage 1 that is most clearly worth describing, as it deviates from the expected normal state and the actual solution is furthermore often unknown. Almost by definition, Stage 1 is *prospective* and Stage 2 *retrospective*, as the situations are only introduced because of each other.

- **Product search**

  Product search involves a Stage 1 in which a product is needed and a Stage 2 in which it has been obtained. In this case it is Stage 2 that is the most interesting in that it deviates from the expected default state and which is therefore worth describing. The product may be described either with a prospective aspect as a type of entity needed by the user or with the default aspect as a product in itself, for instance by referring to the product name.

- **Job search**

In a job search portal, the user is in some initial situation where a job is needed and wants to proceed to a more optimal situation in which a (new) job has been acquired. However, Stage 1 can optionally be described with a retrospective aspect, as seen by the potential future employer. Alternatively, the desired job may be described either prospectively as something needed or in the default aspect.

- **Topical information search**

Topical information searches involve users looking for information on certain subjects. In enterprise search, this can be either because the site in question is of an informational character or because the user wishes to retrieve corporate information. Typically, topical searches target the Subject, but Object-targeting anticipation also occurs.

The following is an example of a Target Analysis as conducted on the Workindenmark log as part of a project aiming to improve the services that help people from other countries find a job in Denmark. Similar analyses have since been made on other Danish job portals, but very little work has been done in other domains, which may therefore present new challenges. The method is still in the process of being perfected and new types of analyses are continually added. Here, only the basics are explained.

10.1.2 Domain-specific assumptions

It is assumed in this analysis that locations within the borders of Denmark, including Greenland and the Faeroe Islands, represent desired work locations, while other geographic references are interpreted as referring to the origin of the user. Most of the latter are names of countries. In a very few cases, users refer to cities outside Denmark. These may possibly represent cases where people want to work for a Danish company while remaining in that city, and thus they may belong in the category of desired work locations. Nevertheless, it was decided that they be assigned to the category of origin, but the number is very small.

References to nationalities and languages have been interpreted as referring to the user's origins, except in queries which explicitly make it clear that languages refer to linguistic skills. This is the case when more than one language is mentioned, when reference is made to teaching or teaching institutions, and when the language in question is no longer spoken, such as *latin*. These latter cases are assigned to the category of competencies.

References to job titles and task descriptions have consistently been interpreted as referring to the desired job rather than a previous job or education. However, engineers' titles have usually been assigned to the category of educational titles along with other academic titles. Whether this was the right choice remains debatable, but a choice had to be made, and it seemed reasonable to consider engineers as basically a class of people with certain educations suitable for a range of different jobs.

Stage 1		Stage 1'		Stage 2'		Stage 2	
Initial state	Interests	Experience	Skills	Need	Methods	Goal	Core
Origin	Situation	Schooling	Field	Locality	Conditions	Workplace	Sector

Fig. 10.1. The 16 basic target classes employed in the analysis of L^{WID}. Each box encompasses a set of four parallel target classes of which the left columns represent locations or states and the right columns domains or situations. While all targets are mental constructs, the top row covers the most psychological targets while the bottom row contains information that is more closely bound to the external circumstances.

10.1.3 Query classification

The initial step of the analysis consists in identifying possible targets on a micro-level, i.e. in this case the things that users of the WID service may reasonably be expected to refer to. The classes employed were as diagrammed in fig. 10.1. The next step is to add necessary subclasses whose members are potentially ambiguous in similar ways, such as nationalities and geographic references as subclasses of Origin. The classes and subclasses used in the analysis of L^{WID} are as follows.

1. S1: Queries describing the user in the current state
    a. Initial state: The state that the user is in at Stage 1
    b. Interests: What the user is interested in (which might, therefore, affect the choice of job); the state of the user's mind at Stage 1
    c. Origin: Where the user comes from
        i. Geographic origin: The user's country, region or town of origin
        ii. Nationality: The user's nationality
    d. Situation: The (psychological) situation in which users find themselves at Stage 1
2. S1': Queries describing users through their job
    a. Experience: Experience that the user has due to earlier work
        i. Certificate: Certificate possessed by the user
        ii. Experience acquired by the user

b. Competencies: The user's skills and abilities

i. Skills: What the user can do; tools that the user can handle

ii. Linguistic skills: Languages mastered[216]

iii. Subject knowledge: Subject on which the user possesses knowledge

The fact that these queries deal with topics that the user knows something about may have contributed to the unexpected distribution in fig. 8.9 where S1' queries seem to pattern not with S1, S2 and S2' queries as expected but rather with T1 and T2 queries[217]. After all, this particular subtype of S1' queries (which makes up 50% of the S1' queries in L^{WID}) would in a different domain have been interpreted as belonging to the T1 type. It is only because they are used with the purpose of describing the user's competencies that they are referred to this type in the context of a job portal.

c. Schooling: Educational background and school of thought

i. Institution: Place where the user was educated

ii. School: School of thought to which the user belongs

d. Field: Field of science or knowledge that the user belongs to

i. Title: Academic title possessed by the user

ii. Academic field: Field of science that the user was educated in

3. S2': Queries describing a desired job through the user

a. Need: What the user needs given present conditions

i. Seniority: The seniority level appropriate to the user

ii. Need type: The type of post (such as job or internship) needed by the user

[216] These were registered separately, but are shown with other skills below because of their insignificant number. This scarcity is in part due to the ambiguity with nationalities, which is also the reason for their separate treatment in the first place.
[217] Of the S2' queries in L^{WID}, 53% refer to geographic locations which typically consist of one word only, thus increasing the difference in length between the S1' and S2' types even further.

b. Methods: The systems and methods that the user would like to apply in future work[218]

    i. System: Systems and tools that the user masters and wants to use in future work

    ii. Method: Methods that the user would like to apply in future work

c. Locality: The place where the user would like to work

d. Conditions: The conditions of work desired and set as necessary conditions for a future job

    i. Working conditions: Working hours, etc.

    ii. Employment conditions: Salary type, etc.

4. S2: Queries describing a job independently of the user

    a. Goal: The actual work that the user wants to do

        i. Job title: The position or type of position desired

        ii. Work task: The kind of work that the user wants to do

    b. Core: What the user wants to work with

        i. Material or product: The material or kind of product that the user wants to work with or produce

        ii. Work focus: The subject or principle that the user wants to work with or promote

    c. Workplace: The kind of place where the user wants to work

        i. Company: The specific company that the user wants to work for

        ii. Environment: The kind of workplace or environment that the user wants to work in

        iii. Company type: The kind of enterprise or organization that the user would like to work for

[218] These are often difficult to distinguish from skills, and hence were only invoked when there were particular reasons to do so.

   d. Sector: The sector or type of cross-sector work that the user wants to work within

      i. Department or activity: The kind of cross-sector department or activity that the user wants to work with

      ii. Area: The sector that the user wants to work in

5. Combining types:

   a. Degree: An academic position desired or a title won by the user (ambiguous)

   b. Training: A position as an assistant desired, or experience as a former assistant (ambiguous)

6. T1: Topical queries

7. T2: Anticipating queries

8. N2: Queries that describe the actions that the user wants performed

   a. Find document: Queries that aim at the retrieval of a specific document

   b. Help me do: Queries which aim at the performing of an action

9. Unclassifiable: Queries which are seemingly without sense or too ambiguous to be classified

Appendix C1 lists their respective frequencies of occurrence, and some examples are given in Appendix C2. The distribution of the main types is shown graphically in fig. 10.2a-b.

The results are highly interesting in themselves, especially when combined with behavioural data such as click and read-through behaviour, which among many other things showed clearly that academics tend to search for their educational titles and quit the search if they do not find anything in that way, presumably because many academic people do not know what jobs to search for. People who enter their nationality and nothing more likewise tend to give up. Both groups of users would thus benefit from inspiration and information. However, the importance of such findings for the construction of the site in question is beside the focus of this thesis and hence will not be discussed in depth here.

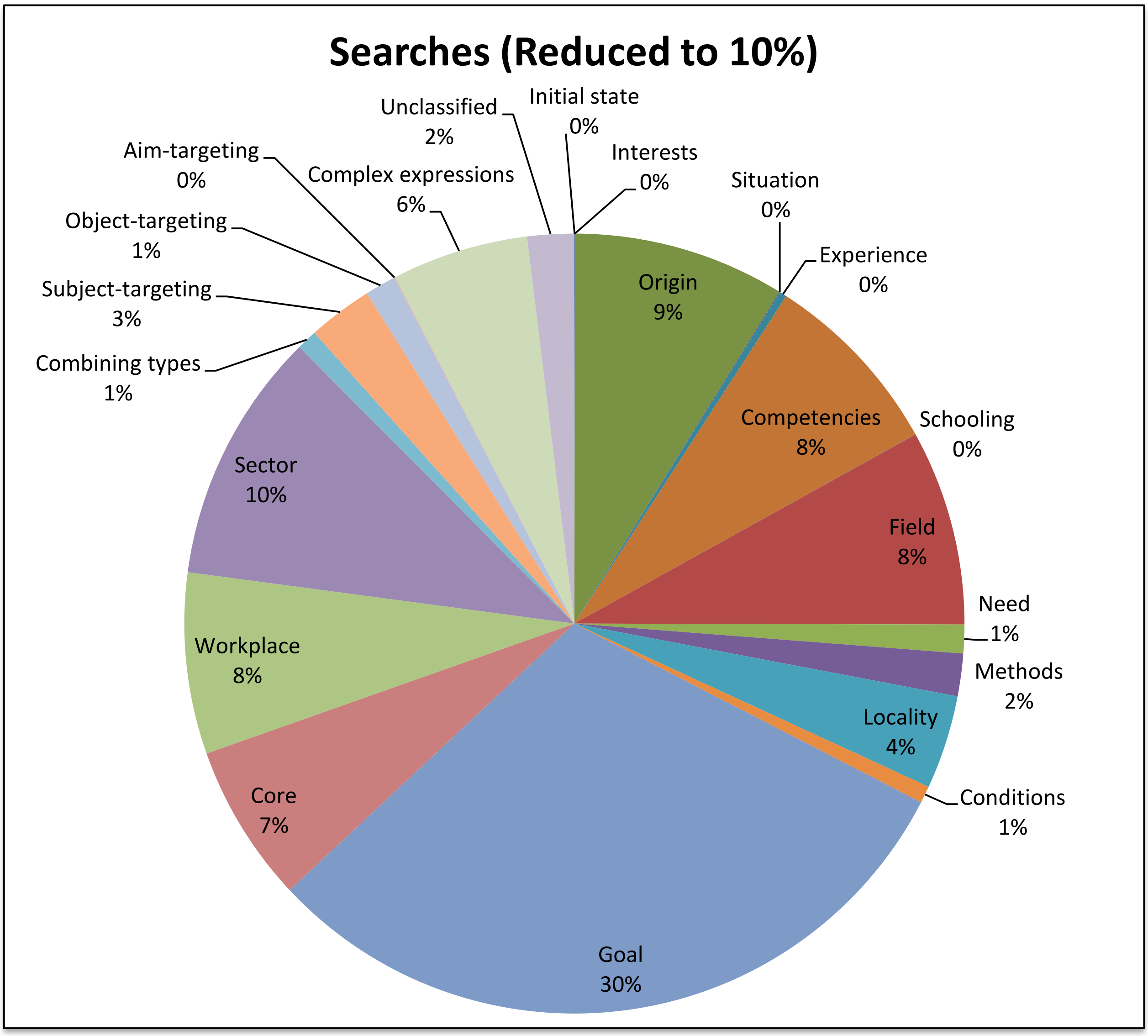


Fig. 10.2a. Results of the Target Analysis of L^{WID}. There are large differences in the frequency of the various targets, but at the same time there is a noticeable spread across target choices. There are clearly many ways of searching for a job.

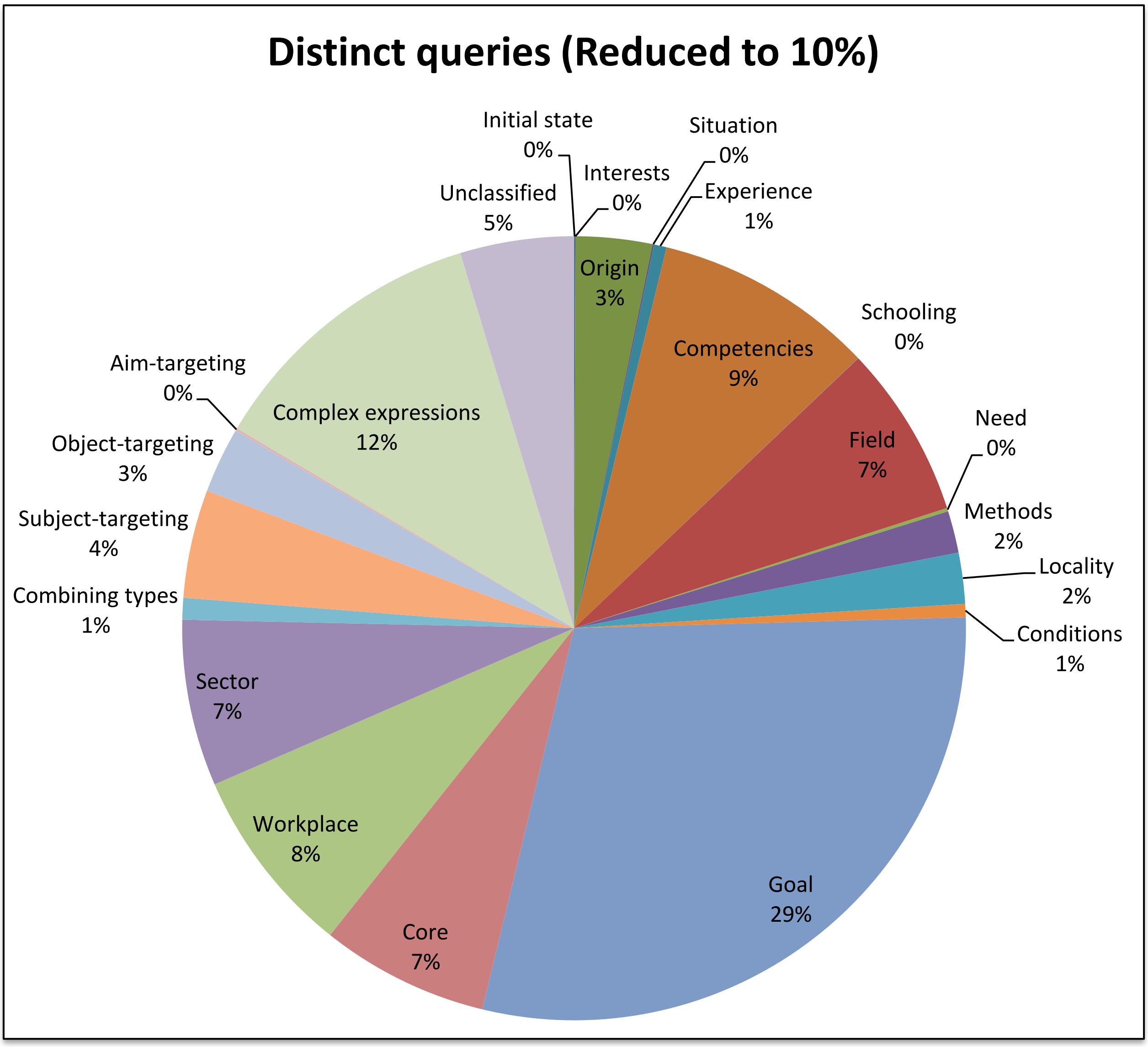


Fig. 10.2b. Results of the Target Analysis of L^{WID}. Note that origins are not very diverse because many users come from the same countries or speak the same languages, while for obvious reasons the complex expressions involving more than one target are relatively rarely used more than once in an identical form. Sectors are also relatively homogenous because many jobs belong to the same sectors.

10.1.4 Results and discussion

Having classified the 9276 distinct queries in L^{WID}, the various classes can now be assigned to targets by comparing with the Stage-based Intent model.

S1-searches describe the user and his or her situation and origin. These queries cover 9% of all searches – most of them by far refer to nationalities.

S1'-searches describe the user's skills and educational background, i.e. they describe the users the way they might imagine that their potential employers would see them. These entries constitute 16% of all searches. Half of these are references to skills and the other half to educational background.

S2'-searches specify what the user is looking for based on his or her needs. These queries may consist in simple terms for work, internships, etc. Alternatively, the reference may be to working conditions, working hours, project basis, etc. Finally, desired geographic locations of the future workplace are assigned to this type, as locations are presumably to a large extent chosen on the basis of where the user intends to be staying, and consequently represent Stage 2 as seen from and determined by the conditions characterizing Stage 1. The geographic searches constitute half of the 8% of all searches belonging to this type.

S2-searches refer directly to the desired job, either through a job title or a description of the desired work tasks. Alternatively, users may specify what they want to work with, what kind of workplace (or even a specific company) the user is interested in, or simply the sector within which they would like to find a job. Not surprisingly, these are common and constitute 55% of all searches; 30% refer to the jobs, 7% to what the users want to work with, 8% to where they want to work, and 10% to the sector.

Some job titles combine several targets so that no distinction can be made. This is the case with *phd*, which is simultaneously a reference to a job title (short for *phd fellow*), which is an S2-search, and a reference to a level of education, which belongs with S1'-searches[219].

3% of the users search for a topic they want to know something about, i.e. they use the T1 target.

1% anticipate strings in documents, including a large amount of people's names. These are presumably to be interpreted as T2-searches.

[219] And of course it may refer to the work product as well.

6% of all searches are complex, i.e. they combine several kinds of references in one query, such as job title and geographic location, or skills and sector.

2% were not classifiable due to excessive ambiguity; one should always expect to find a residual set of unclassifiable queries. Some searches simply cannot be classified and may not even make any sense.

It is remarkable that by far the majority of queries in L^{WID} contain only one target. This indicates that the users are either highly optimistic as to what can be achieved with a very simple search, or they quite simply do not know what to search for. This is perhaps not so surprising, given that the users are generally foreigners who do not speak the language (English and/or Danish), and do not know the Danish society very well either. They consequently need much help and inspiration as regards the jobs that they might look for.

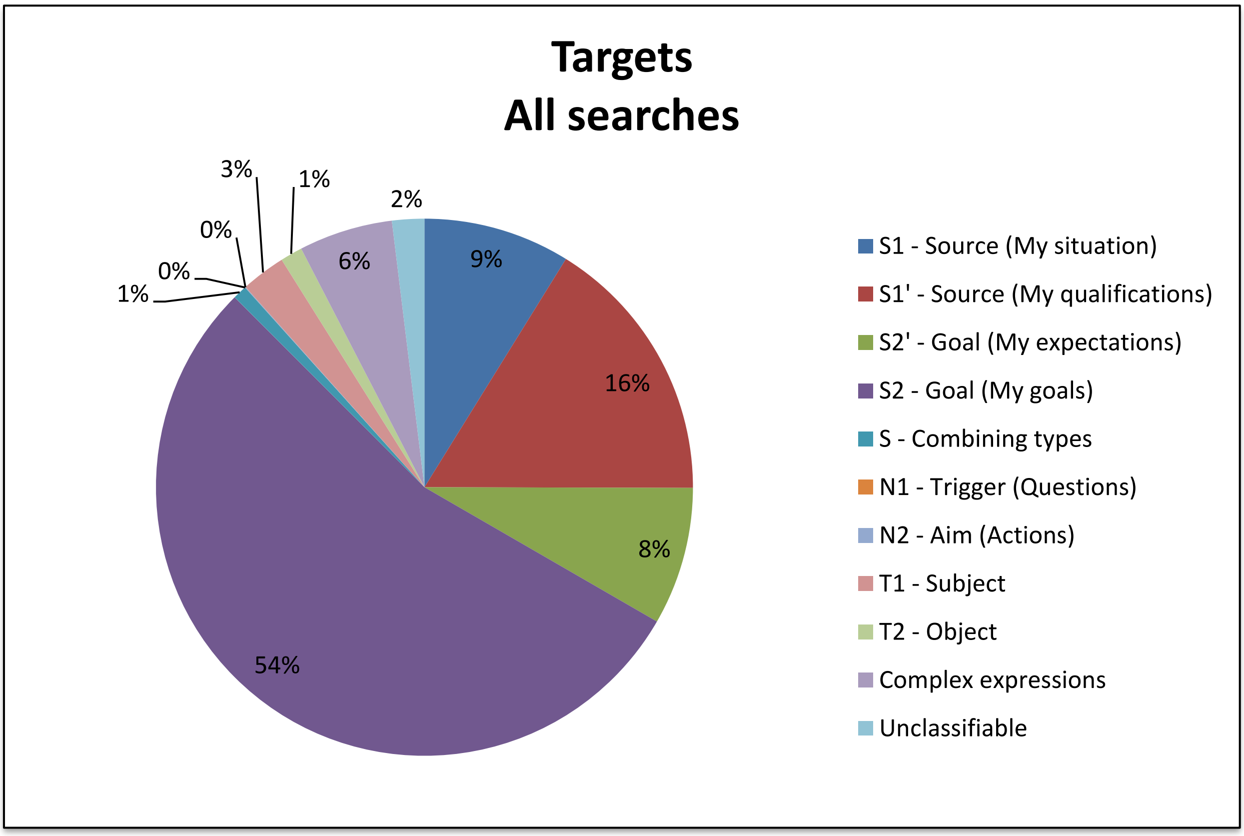


Fig. 10.3a. The use of targets in L^{WID}. The complex expressions mostly (probably even exclusively) belong to the S-dimension, though often across stages. On the other hand, most of the unclassifiable queries could have been assigned to the T2 target, being strings to locate in resource documents.

As might be expected, most users of the WID service target Stage 2 (fig. 10.3a-b). More precisely, if S2' is included, at least 63% of all queries issued refer to this stage, and many complex expressions and many of the unclassifiable queries also belong here, the latter because they include strings which might be interpreted as cases of the Object target. Neither is it unexpected to find that users of the WID service make almost exclusive reference to the S-dimension, i.e. Source and Goal. Among the remaining queries, at least 4% refer to the T-dimension (excluding the possible members deemed to be unclassifiable) while almost none belongs to the N-dimension, facts which are easily accounted for in view of the domain and the purpose of this particular LAT.

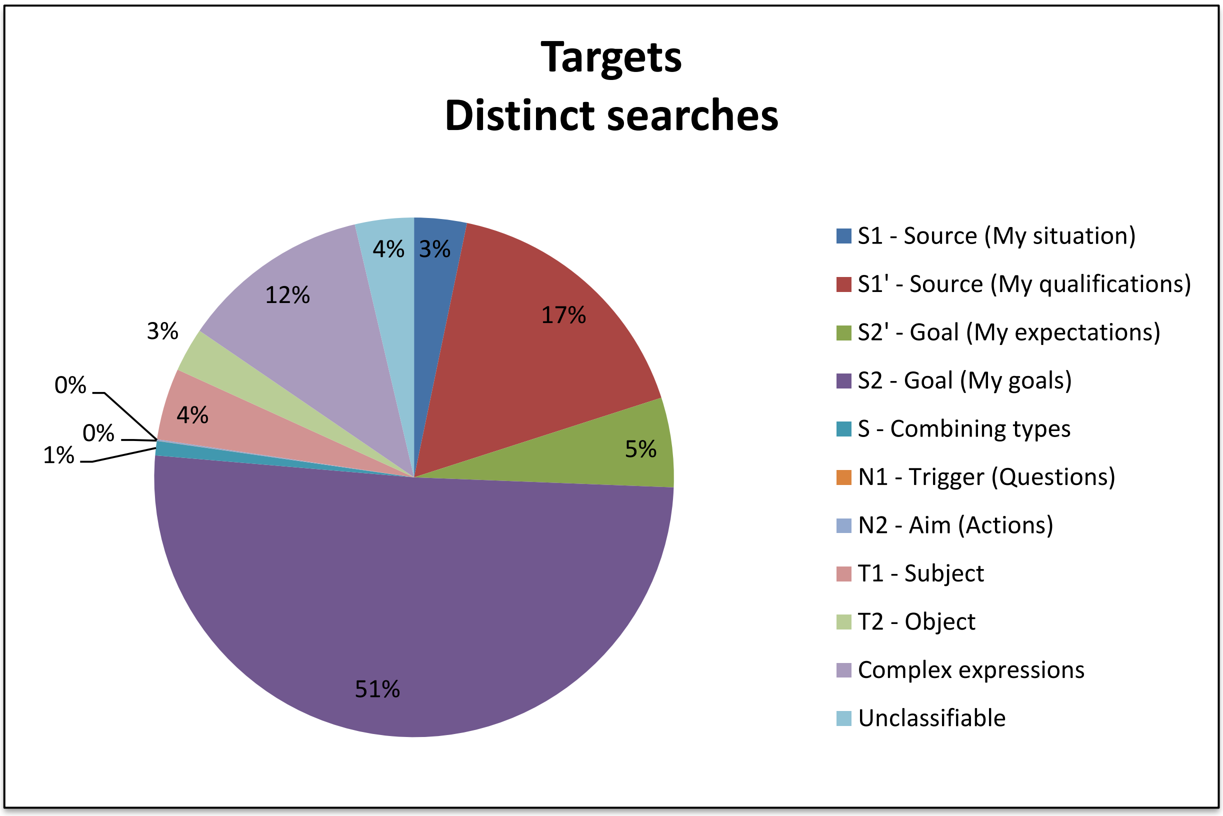


Fig. 10.3b. The distribution of targets in L^{WID}, measured as distinct search strings.

As regards other user behaviour, there are interesting results to be gained too as it turns out that the various target groups represent users with quite different average behaviours. Summing up, and without going too much into detail, four main types of search behaviour can be tentatively observed.

- A tendency to give up after a first attempt without reformulating the query or looking through pages of results (Field)
- A tendency to reformulate (Origin, Skills, Core, Workplace)
- A tendency to look through large amounts of results (Need, Locality, Sector)

- A tendency to succeed in getting results without the need to reformulate or look beyond the first results page (Goal)

This suggests an intimate relationship and interdependency between 1) user behaviour, 2) the kind of target chosen and 3) the user's mental state, with the latter being the initiator of the relationship, though itself affected by the external situation.

The amount of resources a user is willing to look through is an interesting parameter[220]. Jansen *et al.* (1998) found that 58% of their test users went beyond the first 10-result page for more suggestions. The average number of pages viewed was 2.21. Silverstein *et al.* (1999) found that 85% only viewed the first page, with an average of 1.39 pages viewed. However, these results are not necessarily comparable, since these days it is more common for a results page to contain 20 results rather than 10 as in the study by Jansen *et al.*, and if this was the page size in the Silverstein *et al.* study then it is only to be expected that fewer users would find it necessary to go beyond the first page. In an extensive study covering tens of millions of users, Beitzel *et al.* (2007) report that approximately 80% of users viewed only the first page, while no more than 1%-6% went beyond page 2. Unfortunately, like Silverstein *et al.*, they do not specify the size of the pages in their study; it is indeed possible that pages may vary between users. Analysis of L^{WID} shows that in this case users on average view 1.86 results pages, which is a relatively high number due to the domain in question – relatively many users browse through many pages of the results in order to find the best candidate for a job, or several similar candidates[221]. After all, job search is a case where two similar results may often be better than one.

10.2 Search revisited

It has been continuously emphasized throughout this thesis that a query is not merely a string but also a more or less conscious act by the user intended to communicate a need to the system. Search behaviour is determined by human nature and rationality, the expectations that the user has to the requirements and abilities of the system, and the cognitive representation of intentions.

10.2.1 Meta-awareness

The search process starts with a vague sense of motivation and develops through perception of the problematic Source situation and awareness of a kind of need or lack to conscious planning of the search and expression of the need through a query (fig. 10.4).

[220] This parameter is also known as the *futility point* (Blair, 1980)
[221] In the full unreduced WID log the number is slightly lower at 1.77 pages.

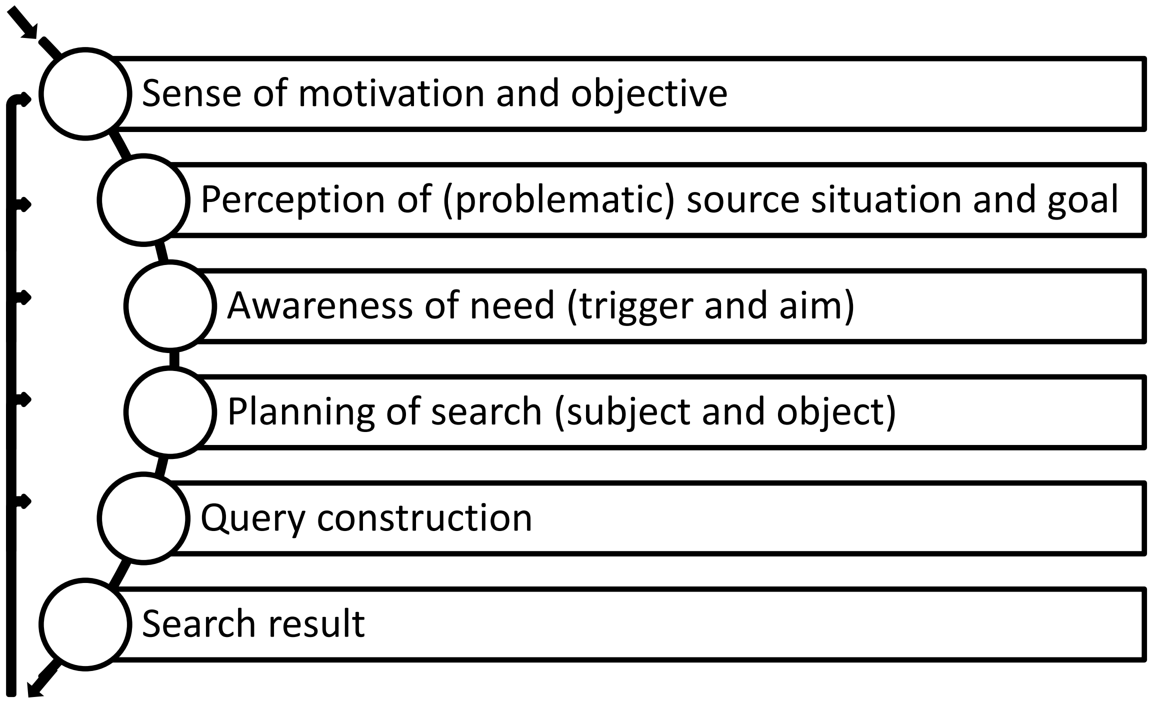


Fig. 10.4. A schematic rendition of the search process.

At any point in this process the user may go back and revise earlier steps, but this is especially common after an unsuccessfully completed cycle which culminated in unsatisfactory results. When this happens, the user is forced to either give up entirely or go back and revise either the expression, the plan, the need (if the results have made the user realize that the need is different from what was thought), the initial problem (i.e., the task), or even the ultimate objective. Before the turn of the millennium, IR system developers had only considered the possibility of revising the query itself, a fact bemoaned by Hearst (1999), who emphasized that searching is a learning process involving revision of needs. Also, goals may be partly fulfilled, whereby a related or partial goal takes precedence. In fact, "users' information needs are not satisfied by a single, final retrieved set of documents, but rather by a series of selections and bits of information found along the way" (Hearst, 1999, p. 264), and hence the process is not just a means to an end, but part of the end itself.

Whatever revision strategy the user adopts (either consciously or unconsciously as a result of being affected psychologically by previous results), not giving up means coming up with yet another expression, which by its very nature may not be the user's best guess or idea of an optimal query (Blair, 2006). This generates further problems, which can only be overcome if the user starts utilizing his or her *meta-awareness* abilities to construct queries that are tailored to the system and document collection in question, as far as their characteristics can be ascertained, which is no easy task. Simon (1976, p.

79ff) summarizes some of the short-comings of such human decision-making behaviour in the following way.

- Rationality requires a complete knowledge and anticipation of the consequences that will follow on each choice. In fact, knowledge of consequences is always fragmentary.
- Since these consequences lie in the future, imagination must supply the lack of experienced feeling in attaching value to them. But values can be only imperfectly anticipated.
- Rationality requires a choice among all possible alternative behaviors. In actual behavior, only a very few of all these possible alternatives ever come to mind.

He also points out that "the word 'all' is used advisedly. It is obviously impossible for the individual to know 'all' his alternatives or 'all' their consequences, and this impossibility is a very important departure of actual behaviour from the model of objective rationality" (Simon H. A., 1976, p. 67).

Unless pressed to do so by unsuccessful results, users do not as a rule disambiguate their queries; the query is just a description of an ideal resource or another target that signals the need, and essentially not a disambiguating expression as regards other, irrelevant resources. Adding discriminators requires the user to switch on an entirely different way of thinking, which operates on a meta-level. Occasionally users do take this meta-level into account when they search. The user might for instance choose to avoid using too vague terms because such terms would result in very low precision and a large recall. Query terms are thus used in two different ways: 1) *descriptors* for describing a target, and 2) *discriminators* for optimizing the distinctiveness of the query. Terms added to a query in order to disambiguate need in principle not occur in the sought documents if the system is able to interpret them as restricting the sense of the disambiguated term. For instance, a query such as *jaguar car* should be able to find documents about jaguars even if they do not mention cars as such, but should exclude resources on the animal. Such a system would of course require the resources to be semantically tagged with the proper sense of the words. In contradistinction to discriminators, descriptors are to be located in documents. Incidentally, the standard way of performing query expansion in order to improve precision relies implicitly on this ability, as it would be common for such systems to add the word *car* to an ambiguous query such as *jaguar* as soon as evidence could be found that this was indeed likely to be the intended sense of the word.

Users have been found to perform relatively poorly when it comes to predicting what terms will trigger the retrieval of the most relevant documents, even though they feel confident about the semantics involved (Ruthven, 2003). In an empirical test it was found that users searching for specific information were only able to retrieve 20% of the documents previously identified as being useful in the collection tested (Blair & Maron, 1985). Furthermore, the participants were convinced that they had found at least 75% of the relevant information. People's meta-awareness abilities are thus rather limited. This, then, is a point where the system would do well to help the user by attempting to counteract its own biases and weak points, as well as those ambiguities which language (and query language in particular) affords. Semantic disambiguation is only the top of the iceberg; semantic relations, syntactic equation (conflating or relating constructions with similar semantic implications), pragmatic interpretation of requests in context and reconstruction of needs and tasks are all part of the job description for a LAT. Much of this is beyond our capabilities at present due to the ambiguities inherent in query expression, but a few openings are identified in this thesis where it would be possible to get a hold. This would provide a foundation on which to build further interpretational capabilities.

10.2.2 The role of a search engine revisited

The question was posed in Part I what the genuine purpose of a search engine is. The reader may have noticed that quite a few definitions have been suggested so far, and it is necessary at this point to establish the relationship between them in order to avoid confusion. All of these definitions have their place in the larger picture of search interaction and can be profitably explained with reference to Saracevic's (1997b) *strata* of information retrieval interaction as detailed below (cf. section 4.2).

On the *situational level*, the user participates in a work task or situation. This level contributes the most general purpose of a search engine, which is the one pointed out by Järvelin (2011).

> The ultimate goal of information retrieval (IR) research is to create ways to support humans to better access information in order to better carry out their (work) tasks.

The way this is achieved on the second level, the *surface level*, where the user interacts with the IR system, is by retrieving and presenting information to the user. This introduces the next level of purposefulness from the user's point of view (translated from Daoud (2009)).

> An information retrieval (IR) system is a system that allows you to find, from a collection of documents, those documents which are likely to be pertinent with regard to an information need of a user, expressed in the form of a query

Or, in the more elaborate form proposed here:

> The task of a search engine is to present the user with any available collection of resources that answer the user's information needs while at the same time (ideally) being highly accessible. The information contained in the documents should also (ideally) be highly accessible and easily applicable to the task at hand.

From the system's point of view this effectively means removing from view all information that is irrelevant, on the *processing level*.

> The task of a search engine is to identify irrelevant information so that it can be hidden from view in order to facilitate navigation among the remaining potentially relevant resources.

So in the end, not just particular search intentions, but even the conception of the overall purpose of the search engine depends on the context that one is focusing on. This emphasizes the fact that context is a crucial part of a query that cannot be neglected, however difficult the incorporation of contextual information is.

On the *cognitive level*, the raison-d'être of an IR system is rather more abstract:

> Search engines should help users find information-as-thing that is likely to trigger information-as-process leading to an expansion of their information-as-knowledge.

Ingwersen (1992, p. 46) stated this in general terms:

> The task of IR is to bring the individual user out of his uncertainty state and close to thrownness, with respect to his *desire for information* by making him obtain information that transforms his knowledge state.

The way to achieve this last goal is by helping the user understand the task and define the information that is needed.

> Search engines should help users understand their needs and interact with the system in ways that secure success in their endeavour to satisfy those needs.

This emphasizes the fact also stressed by Saracevic that the various levels interact and depend on each other. Improvements to a LAT are mainly made on the processing level as algorithms (and the engineering level as regards hardware), on the content level through indexing and meta-tagging and on the surface level in the form of interface design. The choices made here affect the experience on the surface level, which in turn has consequences for the affective and the cognitive level. The latter finally interacts with the situational level and hopefully facilitates the process of finding something useful. Thus, all of these levels are necessary parts of the understanding of search behaviour and all should be optimized so as to ensure maximal user satisfaction and minimize frustration, uncertainty and anxiety.

> Preferably, information retrieval systems should be designed to cope with and support all different kinds of conceivable uncertainties through IR interaction. This implies to help the searcher help the system to support the searcher, etc. – i.e., *making the searcher more informative* towards the system, which, in turn, then increases its informativeness toward the searcher. (Ingwersen & Järvelin, 2005a, p. 300; emphasis original)

Search being a communicative act, responses are evaluated by users according to general processes of relevance judgment. An important consequence of this is that, when searching for a fact, other information than the actual desired answer may at times be more relevant, depending on the situation. This reduces the importance of the user's preconceived needs severely. Rather, relevance is entirely dependent on the user's post-hoc interpretation of the responses provided by the tool. The mission of a LAT is consequently to simulate this interpretation process as if it were in the user's stead in the given situation and present him or her with a reduced set of resources so that the user is spared from evaluating all the resources available in the collection, which would usually be wholly unrealistic. Even so, one should be careful not to fall into *the cognitivistic trap* as Ingwersen (1992) put it; the main interpretation and relevance judgment should be left with the user, as the system will not be better at this than a human anyway. The main task is to reduce the strain on the user by presenting a compact set of results that the user can handle effectively.

10.3 New opportunities

Even though more than 80% of web search queries are said to be informational, existing analyses effectively benefit the other types alone and still treat informational queries as if they were non-cognitive strings.

> With the relatively clear characteristics of navigational and transactional queries, informational queries became the catchall by default. (Jansen, Booth, & Spink, 2008, p. 1262)

Hence, a linguistic analysis of the cognitive reasons for wording a query in a certain way is sorely needed. The attention that is currently paid to linguistic variation in queries is very slight, mostly reduced to accepting morphological variation as an unavoidable nuisance. Referential issues, morphology, syntax, information structure and lexical choice are all crucial for a proper understanding of queries, yet search tools in general largely ignore this kind of information, and even purposefully delete it when it is provided in the form of morphological affixes. It is high time that linguistic information were (re)introduced into IR algorithms. A viable alternative to conflation strategies is to base the search engine on a complete dictionary of actual forms so that the system knows that *dog* and *dogs* are the singular and plural forms of the same word. This is employed in some sophisticated search systems today[222], but there is still some way to go before the insight that forms differ in their meaning in predictable ways can be fully exploited. Query syntax seems strangely enough not to have been investigated much and is a promising field for further study, while stadial variation (*Target Analysis*) is a new type of study entirely, which proves to be both interesting and fruitful.

10.3.1 The challenges

As pointed out by Ingwersen (1992), the perfect semantic representation of queries and documents is not going to solve the problem of retrieving relevant information.

> Such analysis techniques may demonstrate a high retrieval performance ratio – but of what? Of sentences carrying identical, or very similar semantics, to the semantics of the request, but not necessarily carrying *information*. In other words, the searcher may thus retrieve what he already knows from various texts, rather than what he does not know. (Ingwersen, 1992, p. 81)

[222] Such as Ankiro Enterprise Search.

Clearly, if users of a job search portal provide information about their current status or competencies they are *not* interested in information on these topics. Thus, Ingwersen's concern is justified; future IR is not going to be perfect simply by identifying the semantics of query and resource and matching them in a sufficiently fuzzy way. Rather, an interpreting *intermediary* module needs to be added, at the very least in the conceptualization of the IR problem (Ingwersen, 1992) (cf. section 2.4.5). Such an intermediary module would help identify the target and translate the intention into something useful, such as suggesting a job related to the Stage 1 description provided by a user looking for employment.

Ingwersen furthermore advocates an approach called *polyrepresentation*, i.e. representing the information need of a user not just as a (simple) query, but as a combination of several versions of the need (Ingwersen, 1996; Ingwersen & Järvelin, 2005a). For this purpose he suggests "not only the current (often topical) information need, but also (and more importantly) the underlying problem space, actual work task or interest, and the dominant work domain(s)" (Ingwersen, 1996, p. 41). This idea is similar to the notion of a *complex query* suggested here. Of course, ideally it would seem that not three but eight different versions would be needed (corresponding to the eight targets), though they could hardly all be expected to contain information in every case, since obtaining any information at all apart from the literal query is hard enough, never mind aiming at the full range of potential targets. Nevertheless, the Stage-based Intent model provides a schema of variables that can be filled in as the intermediary learns about the user in a given session.

Improved natural language processing of queries is a prerequisite for such intent decoding.

> [O]ne thing seems certain for the future: NLP and other linguistic resources will become – if they are not becoming already – indispensable parts of every effective IR system. (Arampatzis, van der Weide, van Bommel, & Koster, 2000)

The classification of queries according to user intentions presupposes an understanding of intentions and a typology of intention types. This is lacking in existing accounts; users have intentions on multiple levels, often in parallel, and query classification is useless if this complexity is not recognized. Even semantic search engines currently perform their linguistic analyses on the substantial rather than the formal query. This is certainly a huge step forward from the situation in which there is no linguistic analysis at all, but there is still far to go. Human addressees are able to reconstruct the linguistic message behind an utterance even though they do not have access to it either, so in principle at least it should be possible for LATs to achieve this too.

We saw in Part II that topicality is not just a matter of interpreting the semantics of a topic label either; there are different kinds of topicality. There is a huge difference between a system that attempts to give users what they ask for (*focal topicality*) and one that is constructed to understand the Source situation (*local topicality*). The latter kind of system would often seem to ignore the user's wishes, and would hence potentially generate a lot of dissatisfaction initially, but in the end it would attempt to satisfy users even more than by simply returning what they ask it to. If users knew that they were communicating with such a system, it might potentially be very helpful in some cases, but in most instances they probably prefer systems that do what they tell them to, even if that leads them through a lot of dead ends.

In the domain of job search, Ankiro One-Line Search is able to identify the query *linguist Copenhagen* as consisting of a job title followed by a location, but whether this was meant to be a Stage 1 or Stage 2 query cannot possibly be determined from the query alone. If the information is considered requisite, the system might be constructed so as to ask the user whether he or she *is* a linguist from Copenhagen or is looking for a job as one. Sometimes the difference may indeed be of considerable importance, yet in other cases it can be ignored and the question would merely annoy the user. A good solution would probably be to retrieve all results immediately and allow the user to post-filter results or answer questions for improved results when necessary. A viable alternative is to present the different responses side by side.

10.3.2 A new framework

In a review of IR research, Ingwersen and Järvelin (2005a) make the following pronouncement about the state of the art.

> [T]here are still no frameworks that are broad – from work task to information retrieval – abstract – specifying theoretical concepts and their relationships rather than concrete stakeholders – analytical – suggesting hypotheses – and explicitly process-oriented. Such frameworks remain to be created and empirically validated, first in specific domains and contexts, and then more generally. (...) There is too little theory-driven research although recent work suggests that there is progress towards this goal. (Ingwersen & Järvelin, 2005a, p. 105)

It is hoped that this thesis will be considered to contribute to such progress in the field, as did their own work. We have presented above a framework which is able to model users' expressive choices from their first motivation to the planned search interaction and the ensuing relevance assessment. This satisfies several of the pressing needs in the IR research field. The suggested framework deals specifically with query expression rather than general user behaviour. It therefore lacks (or rather, does not explicitly treat) several elements associated with the cognitive framework such as laid out by Ingwersen and Järvelin (2005a). The reader is referred to that book for more about such issues. Ingwersen and Järvelin's model is even more inclusive than the one suggested here (for which they deserve commendation), and thus the suggested framework fits into their model, but their focus is on external circumstances rather than the intimate issues covered by the models proposed here, which is why a new framework is needed for this part of the model. Information on work tasks and the user's knowledge level and psychological development etc. can be easily fitted into the framework suggested here when needed; this has simply not been our priority, because while certainly contributing to the investigation of interesting user behaviour, the strong focus on such behavioural traits is exactly what makes existing models inadequate when it comes to explaining query form. In particular, while Ingwersen and Järvelin (2005a, p. 306) focus on contextual features which "directly influence the involved actors' perception of the situation at hand" because their main interest is with need development in a situational context and finding out what the user's real needs are[223], the point of view chosen here causes us to be primarily interested in that perception itself rather than what formed it. From this point of view, the perception that the user has of the situation *is* the situation, and this acts as context for the expression of a query. As Ingwersen and Järvelin themselves point out in response to a potential objection against excluding even more remote societal politico-economic context, "The key is to be able to *capture analysable evidence*" (2005a, p. 376; emphasis original). Obviously, both kinds of investigation are needed in order to satisfy users' needs, and I am certainly not claiming to be uninterested in issues relating to task perception, but the query is, after all, the main source of data available to the system, and so uncovering the reasons for wording a query in a particular way is of the utmost importance.

The Stage-based Intent model complements and fits beautifully into the Faceted Stimulus model suggested in Part I and the Regulated Flux model of relevance presented in Part II. Intuitively, it should be possible to predict the relevance of a resource on the basis of the possible targets, as long as they are known in sufficient detail. However, the Stage-based Intent model does not incorporate any information on the actual topic or type of resolution sought; it only reports what type of response is desired. It deliberately avoids any reference to the semantic contents of the query terms, firstly because they are so often ambiguous, and secondly because they are in general handled more or less

[223] Cf. "we are interested in a cognitive, task-based perspective on information seeking and retrieval" (Ingwersen & Järvelin, 2005a, p. 313).

successfully by other means. The actual values such as descriptions of Source situations, topic domains, etc., represent contextual information, and as such they are covered by the Facet-based Context model developed in Part I, which also connects directly to the Faceted Stimulus model and through it to the Regulated Flux model. Together, these four models make up one coherent super-model of search reasoning and reaction[224]. The models await validation through further experiments and data analysis, and many of the behavioural claims made likewise need to be thoroughly documented by psychological experiments. Such proofs are beyond the scope of this thesis, where it is assumed, for instance, that meta-awareness is an extra add-on that needs to be triggered by attentive users and that the neutral state of affairs is for search to be carried out in the same way as discourse, i.e. by drawing heavily on context in the construction of a message. Much interesting work remains to be done in confirming such hypotheses. At present even a little increase in our understanding of these issues represents significant progress. In fact, one of the main conclusions to be drawn is the insight that search is even more complex and difficult to grasp than previously thought. The Faceted Stimulus model, the Regulated Flux model and the Stage-based Intent model together demonstrate how much is going on with very few clues given to an interpreting system. The best we can do is probably to collect the necessary information directly from users themselves, but even then we shall have to contend with the problem that they are rarely aware of their reasonings and the effects of their emotional reactions.

Along the way we have found it necessary to define a range of notions such as information, relevance and cognition, all of which in many traditions have narrower definitions, but where authors differ in their delimitations so that only the most general definition is able to encompass them all. I do not consider the narrower definitions found elsewhere as problematic in principle as long as people do not insist on such a definition in general; it is perfectly legitimate to adopt a narrow definition for a particular purpose. However, when discussing the phenomena for their own sake and working towards an understanding of them, it is in all cases the general version that is the most revealing, because with a broad definition it is possible to discuss the internal structure of the concept and consider the relationships between subtypes. Hence, I have consistently advocated such general understandings.

10.3.3 Validity and future work

Most queries may in principle be considered string searches and thus to target the Object, but this analysis hardly helps us. Instead, we need to explore what other interpretations suggest themselves in addition. Likewise, most queries may in principle be interpreted as topic searches targeting the Subject.

[224] I consciously avoid calling it a model of search *behaviour*, because we have not attempted to model interactive behaviour in detail the way that Kuhlthau (1993) did, among many others.

However, once again we can only make progress if we attempt to classify as many queries as possible as not belonging to this target, because that tells us much more about the contextual need of the user. We cannot help people much with their string searches, and the day may be approaching when the same is true of topic searches, if semantic indexation becomes sufficiently improved to make identification of topics reliable. Then, it will be clear that identifying topics is not the solution to all of the users' problems; the only way forward will then be to identify the Triggers, Sources and Motivations etc. behind the searches in order to establish what the user really needs.

Throughout this part of the thesis, one of the main goals has been to devise a method of classifying queries based on features of the expression only. A complex model has been developed which is thought to be closely related to the psychological process underlying search. However, strictly speaking the ultimate unbiased classification has in the end turned out to be an impossibility, partly because language is under-differentiated and partly because the facets of queries are an inherent feature of linguistic expressions and hence mostly not recoverable from them. In other words, we cannot know whether a certain user intended a noun to be interpreted as a linguistic string, a topic or a partial description of a situation. In fact, it would often have more than one of these interpretations simultaneously. Even so, it is possible to classify a great many queries according to the target types, and even occasionally according to facets. One must simply acknowledge that certain pairs of types are often indistinguishable and classify queries into specifically constructed ambiguous types accordingly. The variation is certainly important to capture, as a classification algorithm which looked at the substantial queries only would class *all* queries as targeting the Object, while one that went no further than the representational facet would classify most queries as Subject-targeting, i.e. as topical queries, which is in fact what has been the reality of IR studies for a very long time. We ought now to take the next step and include the interpersonal facet and allow the remaining six targets to enter the stage. How this is to be done is a matter for future work; suggestions would be paying attention to the Source situation and allowing users to search based on the characteristics of their problems when this is helpful, or based on Motivations such as a need to be entertained, Triggers if they need answers to questions or facts, and Aims when they target actions that they want performed. Much of this has of course been attempted before, sometimes successfully, but the full picture has not been available, and hence some users have always been left in the cold. Mostly, these techniques have been confined to specialized search systems that only help people with questions or those with certain types of problems and so on, whereas the suggestion made here is that these features ought to be available in standard enterprise search as options for alternative search whenever a traditional search fails to be helpful.

The difficulty of decoding intent is not a shortcoming of the model; it is a fact of reality. Apart from forwarding the understanding of search, the framework is useful because it can be used for preliminary manual and to some extent semi-automatic analysis, followed by more automatic processes when the domain and the search behaviour occasioned by it are properly understood. Many choices that from a general-context point of view might seem arbitrary must be made in order for the method to work in any particular domain.

There are several types of queries whose classification with the suggested framework is debatable. It may be discussed, for example, whether a particular topical term targets the Subject or the representational facet of the Object, or whether a query targets the Aim or the Trigger if it refers to a certain type of resource but the returned resource was known in advance. It is not always easy to come up with an answer to such questions, but it is likely that some of them will be settled in future. The very fact that such discussions can be had is proof of the success of the framework, which was introduced exactly for this purpose.

Conclusion

Query transaction logs are in general considered to be useful mostly for analysing large amounts of data with statistical methods; it is generally believed that it is not possible to decode users' intentions from queries alone. However, given that they are linguistic messages from users to the system, queries do actually convey intentions and so it should be possible in principle to extract such information. What this kind of analysis requires – apart from access to contextual data – is a framework for understanding intentions and the choices that users have when wording their queries, and a method for analysing queries with reference to the contextual part of the message which is implied but not explicitly expressed.

With a view to improving transaction log analysis, and through that also the IR systems themselves, we accordingly set out to develop a set of models which would help understand why users come to express their needs in particular ways. Developing the framework necessitated discussion of context, relevance, need development, and the cybernetics of search, all of which are controversial topics. We achieved our goal with a framework consisting in the main of four models, all of which relate to each other in intimate ways, along with a formal notation of queries.

The *Stage-based Intent model* combines what we know about the psychology of the search process with data from query logs and visualizes the choices that users have when deciding how to word their queries. The model consists of eight referential *targets* that users may choose to refer to with a given query term.

Four of the targets furthermore have six (or eight) *facets*, and a query term may target any one of these, depending on whether a string, a word, a topic, some helpful information, etc., is sought. The *Faceted Stimulus model* deals with the faceted nature of query terms, topics, resource documents, and task situations.

When resources are evaluated for relevance, these same facets are at the heart of the process, and evaluation occurs along six (or eight) corresponding *dimensions*, each measurable *bisubjectively* along three scales. The *Regulated Flux model* clarifies the relevance assessment process and regulates what features of a resource are assessed at a given point in the process.

In order to simulate relevance evaluation, contextual information must be collected, and for this purpose the *Facet-based Context model* was developed in order to clarify what parts of context are relevant to a search.

Finally, we devised a method of analysing query logs informed by the above models called *Target Analysis*. This method consists in modelling the domain according to relevant target types and classifying queries with rigorous reference to query form and careful observance of ambiguity. The results of a sample analysis were presented and were shown to be highly informative about the uses of the tool in question. Among other things, it confirmed the feeling underlying the models that the uses of a site are manifold and that, far from being mainly topical, queries belong to a multitude of different types which require separate treatment.

Appendices

Appendix A – An example of a facet-based context model

For reference, a preliminary attempt at a full model with suggested contextual data of potential relevance to IR is shown below. The actual contextual data types suggested below are just examples; future research may result in an improved list. Also, one should keep in mind that this is a model of all potential contextual information relating to the search and not a recommendation on what information to gather in an actual search system. Further research may establish a hierarchy of importance among the variables so that the number may be reduced to a more manageable set, possibly depending on the type of search.

1. The intentional task

This refers to the task as a moral plight.

a. *Task environment*: The background of the user, including long-term tasks. The user's moral views.
b. *Consequences of task environment*: User type.
c. *Task structure*: The user's responsibilities and plights.
d. *Task history*: Moral developments and changes in responsibilities.
e. *Consequences of task structure and history*: Search type.

2. The substantial task

This refers to the task confronting the user including the physical user him- or herself, as perceived by the user.

a. *Task environment*: The spatiotemporal context of the task as perceived by the user. Current time and location of the physical user at the time of search may be used as proxies unless otherwise stated, though they need in no way coincide.
b. *Consequences of task environment*: User nationality.
c. *Task structure*: The task as perceived by the user. The state the physical user is in when approaching the LAT. This might be useful in a health care search, for instance. User age and sex.
d. *Task history*: Previous situations which have affected the physical user.
e. *Consequences of task structure and history*: The user's legal rights, depending on age, etc.

3. The formal task

This refers to the task as a situation in which the user must find his or her place; something to be analyzed and handled.

a. *Task environment*: Entities involved. Relationships to other entities.
b. *Consequences of task environment*: Types of entities involved. User’s profession.
c. *Task structure*: The task as conceived by the user. The mental state the user is in when approaching the LAT, such as anxiety or frustration leading to a need for a response. The user’s age, language and gender. Note that such parameters are of importance to various layers of the user, both physical and cognitive.
d. *Task history*: Previous situations which have affected the user. Documents previously consulted by the user.
e. *Consequences of task structure and history*: Cognitive context such as the user’s level of expertise and demands as regards resources. Also, the stock of knowledge acquired from documents previously consulted by the user.
f. *Task-user relationship*: The user’s emotions as regards the task.
g. *Task-user novelty*: Novelty of a task to the user.

4. The representational task

This refers to the task as being of a certain type and analyzed by the user for causes and effects.

a. *Task environment*: Overall domain to which the task belongs, type of task. Relationships to other problems, tasks and workflows.
b. *Consequences of task environment*: The user’s role in the task. The complexity of the task.
c. *Task structure*: The task as understood by the user.
d. *Task history*: Previous tasks confronted by the user.
e. *Consequences of task structure and history*: User type.
f. *Task-user relationship*: The user’s interests and preferences in general.
g. *Task-user novelty*: Novelty of a task to the user.

5. The interpersonal task

This refers to the task as a problem to be solved.

a. *Task environment*: Possible solutions.
b. *Consequences of task environment*: Possible outcomes.
c. *Task structure*: The task as interpreted by the user.
d. *Task history*: Stages in a work process, and search steps related to the intention of achieving a result.
e. *Consequences of task structure and history*: Results achieved so far, potential gains.
f. *Task-user relationship*: The user’s general aims and intentions.
g. *Task-user novelty*: Novelty of a problem, task or workflow to the user.

6. The effective task

This refers to the task as something that affects the user psychologically.

a. *Task environment*: Feelings towards the search task.
b. *Consequences of task environment*: Emotional state of the user.
c. *Task structure*: Feelings regarding the work task.
d. *Task history*: Emotional developments.
e. *Consequences of task structure and history*: Emotional state.

7. The interpersonal query

This refers to the query as a message from the user to the system.

a. *Query environment*: The search task environment.
b. *Consequences of query environment*: Mode of searching.
c. *Query structure*: Type of query.
d. *Query history*: History of requests made by the user.

8. The representational query

This refers to the query as a description.

a. *Query environment*: Overall domain to which the user's needs belong, type of task. Relationships to other problems, tasks and workflows.
b. *Consequences of query environment*: The complexity of the task.
c. *Query structure*: The task as portrayed by the user.
d. *Query history*: Stages in the expression of a need. Search steps related to the expression of the need.

9. The formal query

This refers to the query as a linguistic expression.

a. *Query environment*: Conceptual level of used terms.
b. *Consequences of query environment*: Mode of expression.
c. *Query structure*: The linguistic structure of a query.
d. *Query history*: Stages in the expression of a need. Search steps related to the expression of the need.
e. *Consequences of query structure and history*: Wordiness, expressive organisation and linguistic presentation of the need.
f. *Query-LAT novelty*: Novelty of the employed concepts to the system.

10. The substantial query

This refers to the substantial query as issued by the user, including any additional information provided to the system.

a. *Query environment*: The type of query. This is always a text string in connection with LATs, by definition, though in a broader study it might alternatively be activated by, say, an auditory stimulus, a scanned picture, iris recognition, or a pin code, etc.
b. *Consequences of query environment*: The substantial format of the query.
c. *Query structure*: The actual query.
d. *Query history*: The stage reached in the overarching search process. Frequently used queries.
e. *Query-LAT relationship*: Other people who used the same or similar queries.
f. *Query-LAT novelty*: Novelty of the query terms to the system.

11. The intentional resource

This refers to the resource as intended for someone.

a. *Resource environment*: Uses of the resource.
b. *Consequences of resource environment*: The feeling that the resource ought to be accessed.
c. *Resource structure*: An imposed need to access the resource.
d. *Resource history*: Changes in the necessity to access the resource.
e. *Consequences of resource structure and history*: The need to access the resource.
f. *Resource-user relationship*: The feeling of a need to access the resource.
g. *Resource-user novelty*: Novelty of the need to access the resource.

12. The substantial resource

This refers to the substantial resource as provided by the author to the system, including any additional information included as metadata.

a. *Resource environment*, i.e. the device and website or application from which the search is performed and the resources accessed. Also, the spatiotemporal context of the device, and hence the user, at the time of search. This properly belongs with the device and not the user, because it may be irrelevant to the user's needs if the search is performed at a different time and place from the task. However, it is highly useful in ubiquitous computing with mobile devices.
b. *Consequences of resource environment*: The format of the resource.
c. *Resource structure*: The type of resource - Text, image, video, audio, etc.
d. *Resource history*: Alterations made to the substantial document. This belongs with the substantial document; contents cannot be altered, as they are constructed anew at every reading. Click-through behaviour, favourite resources, etc.

e. *Consequences of resource structure and history*: Layout clarity, aesthetic appeal. Suitability for the blind, deaf and so on. Organisation and presentation of the contents, layout. Size of resource.
f. *Resource-user relationship*: Availability of the resource. This includes the cost (in money and time) of obtaining substantial resources.
g. *Resource-user novelty*: Novelty of the substantial document to the user. This is zero in navigational searches where the user merely wants to be brought to a certain page that he or she already knows.

13. The formal resource

This refers to the resource as a linguistic expression.

a. *Resource environment*: The linguistic context of a retrieved passage.
b. *Consequences of resource environment*: Mode of expression.
c. *Resource structure*: The linguistic structure of a passage.
d. *Resource history*: Alterations made to the linguistic document.
e. *Consequences of resource structure and history*: Readability, linguistic simplicity and level of linguistic difficulty. Wordiness, expressive organisation and linguistic presentation of the contents.
f. *Resource-user relationship*: The availability of the time required to read the linguistic document.
g. *Resource-user novelty*: Novelty of the linguistic document to the user.

14. The representational resource

This refers to the resource as a semantic text with portrayals of situations.

a. *Resource environment*: Overall domain to which the topics of the resource belong. Relationships to other concepts, topics and domains.
b. *Consequences of resource environment*: The generality level of the resource domain.
c. *Resource structure*: Semantic content and topics of the document.
d. *Resource history*: Stages of investigation of the portrayed topic, popularity in media.
e. *Consequences of resource structure and history*: Interestingness, entertainment value, and amount of similar semantic information available from other resources.
f. *Resource-user relationship*: The user's main interests and preferred topics.
g. *Resource-user novelty*: Novelty of a topic or concept to the user.

15. The interpersonal resource

This refers to the resource as a message from an author to a reader.

a. *Resource environment*: Relationships to other resources. Resource type.
b. *Consequences of resource environment*: The complexity of the resource.
c. *Resource structure*: Interpersonal contents of the resource.
d. *Resource history*: Stages of campaigning, publicity in media.
e. *Consequences of resource structure and history*: Informativity of contents, resulting cognitive state, and level of expertise gained. Interpersonal clarity. Amount of similar interpersonal information available from other resources.
f. *Resource-user relationship*: The user's interests, preferences and emotions with regard to this topic.
g. *Resource-user novelty*: Novelty of a message to the user.

16. The effective resource

This refers to the resource as something with a psychological effect on the recipient.

a. *Resource environment*: Relationships to other types of resources. Effect type.
b. *Consequences of resource environment*: The complexity of the resource.
c. *Resource structure*: Emotional effects of the resource.
d. *Resource history*: Emotional developments during the reading of the resource.
e. *Consequences of resource structure and history*: Emotional state of the reader.
f. *Resource-user relationship*: The user's interests, preferences and emotions with regard to certain effects and manipulations.
g. *Resource-user novelty*: Novelty of an effect to the user.

17. The effective cause

This refers to the cause as something which affects the author emotionally.

a. *Cause environment*: Feelings towards the job of campaigning.
b. *Consequences of cause environment*: Emotional state of the author.
c. *Cause structure*: Feelings regarding the cause.
d. *Cause history*: Emotional developments.
e. *Consequences of cause structure and history*: Emotional state.

18. The interpersonal cause

This refers to the cause as a project for the author.

a. *Cause environment*: Relationships to other causes. Cause type.
b. *Consequences of cause environment*: The complexity of the cause.
c. *Cause structure*: The cause behind the production of the resource.
d. *Cause history*: Stages of campaigning.
e. *Consequences of cause structure and history*: Publicity in media.
f. *Cause-author relationship*: The author’s intentions and interests in general.
g. *Cause-author novelty*: Novelty of a cause to the author. Author competency.

19. The representational cause

This refers to the cause as a being of a certain type.

a. *Cause environment*: Overall domain to which the cause belongs, type of cause. Relationships to other information campaigns.
b. *Consequences of cause environment*: The author’s role in the cause. The complexity of the cause.
c. *Cause structure*: The cause as understood by the author.
d. *Cause history*: Previous causes taken up by the author.
e. *Consequences of cause structure and history*: Author type.
f. *Cause-author relationship*: The author’s interests and preferences in general.
g. *Cause-author novelty*: Novelty of a phenomenon to the author. Author competency.

20. The formal cause

This refers to the cause as something the author must handle.

a. *Cause environment*: Entities involved. Relationships to other entities.
b. *Consequences of cause environment*: Types of entities involved. Author’s profession.
c. *Cause structure*: The cause as conceived by the author. The mental state the conceiving author is in when deciding to write, such as anger or frustration leading to a desire to communicate. The author’s age, language and gender.
d. *Cause history*: Previous situations which have affected the conceiving author. Documents previously written by the author.
e. *Consequences of cause structure and history*: Cognitive context such as the author’s level of expertise.
f. *Cause-author relationship*: The author’s emotions as regards the cause.
g. *Cause-author novelty*: Novelty of a concept to the author. Author’s knowledge.

21. The substantial cause

This refers to the cause confronting the author, including the physical author him- or herself.

a. *Cause environment*: The spatiotemporal context of the cause as perceived by the author. Current time and location of the physical author at the time of writing may be used as proxies unless otherwise stated, though they need in no way coincide.
b. *Consequences of cause environment*: Author nationality.
c. *Cause structure*: The cause as perceived by the user. The state the physical author is in when deciding to write. Author age and sex.
d. *Cause history*: Previous situations which have affected the physical author.
e. *Consequences of cause structure and history*: The author's legal rights, depending on age, etc.
f. *Author-user relationship*: Whether the user knows the author personally.
g. *Author-user novelty*: Novelty of author or journal, etc., to the user.

22. The intentional cause

This refers to the cause as a moral plight.

a. *Cause environment*: Social and political climate. Relationships to other causes.
b. *Consequences of cause environment*: The author's plight.
c. *Cause structure*: The author's moral views.
d. *Cause history*: Moral developments in society.
e. *Consequences of cause structure and history*: The author's political identity and affinity.

Appendix B – A formal representation of queries

The formal representation of queries developed for this thesis facilitates discussions of query classification. It is based on the suggested Stage-based Intent model. There are four layers of notation.

- **Morphosyntax**

The innermost layer is the morphosyntactic form. Not all details are represented, as only certain characteristics of morphosyntax are salient in search. Noun phrases are not pulled apart, as the entire phrase has been borrowed from natural language and hence nothing can be said about it that does not also apply to natural language. Hence, entire noun phrases are marked $_{n.sg}$ or $_{n.pl}$ depending on the number, or $_{n.prop}$ if it is a proper noun. Predicative adjectives are treated as noun phrases, as are all adjectives not accompanied by a noun. This is because each phrase of the form (A) (N), where either the noun or the adjective can be left out, has a special function in search. As regards verbs, they are marked according to their morphological form, including nominalizations ($_{v.nom}$), which are treated as verb forms due to their functions in search, where they often replace other verb forms and pattern with infinitives.

- **Information structure**

The information structure of the query as coded in its syntax is represented formally with square and round brackets. Square brackets indicate concepts and their sub-concepts, and round ones indicate topics and their sub-topics. A broader concept which has been added after the main concept for disambiguation is marked with < between the concept, and a narrowing modification may optionally be marked with >.

- **Target structure**

Curly brackets indicate the limits of target entities, i.e. the part that refers to a single target. A query that refers to multiple targets thus contains several sets of curly brackets. Within the brackets the type of target is noted in abbreviated form, followed by a colon. The abbreviations are as follows.

C1	Stage 1 cause: Motivation
S1	Stage 1 state: Source
N1	Stage 1 need: Trigger
T1	Stage 1 topic: Subject
T2	Stage 2 topic: Object
N2	Stage 2 need: Aim
S2	Stage 2 state: Goal
C2	Stage 2 cause: Objective

A ‘ may be added to the designation S1 to indicate retrospective aspect and to S2 to signal prospective aspect.

● **Metastructure**

Operators and other meta-communicative items are enclosed in pointy brackets: <and>.

Square brackets may be left out if they coincide with round ones, and round brackets may be left out if they coincide with curly brackets. This reduces the number of brackets significantly and makes reading and writing the formal notation quick and easy.

Appendix C1 – Target distribution in L^{WID}

This table contains data on the distribution of microtargets in L^{WID}.

Target	Micro-target class	Subclass	Searches (x 10)	Distinct queries
S1	Initial state	Initial state	5,0	5
	Interests	Interests	0,0	0
	Origin	Geographic origin	273,7	129
		Nationality	1502,4	164
	Situation	Situation	5,0	5
S1’	Experience	Certificate	47,5	37
		Experience	13,9	12
	Competencies	Skills (incl. linguistic)	1196,7	514
		Subject knowledge	379,8	331
	Schooling	Institution	1,0	1
		School	0,0	0
	Field	Title	919,3	419
		Academic field	718,0	242
S2’	Need	Seniority	74,0	13
		Need type	166,4	95
	Methods	System	307,4	111
		Method	49,2	49
	Locality	Locality	785,4	195
	Conditions	Working conditions	136,5	47
		Employment conditions	5,6	4
S2	Goal	Job title	5209,8	2277
		Work task	945,9	432
	Core	Material or product	1083,2	543
		Work focus	245,2	100
	Workplace	Company	609,5	368
		Environment	882,1	343
		Company type	27,3	11
	Sector	Department or activity	551,0	135
		Area	1560,0	502
S1’/S2’	Combining types	Degree	107,7	20
		Training	63,1	63

T1	Subject-targeting	Tell me about	542,5	414
T2	Object-targeting	Find text	258,8	251
N2	Aim-targeting	Find document	2,0	2
		Help me do	8,0	8
	Complex expressions	Complex expressions	1143,2	1094
	Unclassifiable	Unclassifiable	393,0	340

Appendix C2 – Examples from the Target Analysis of L^{WID}

This appendix contains some examples of queries of various kinds in L^{WID}. The queries shown have not been selected but simply represent the 25 most common ones and where applicable also 25 queries with minimum frequency.

Initial state

The query describes the user before the search, irrespective of the problem that triggered it.

Query	Frequency
enkemand 'widower'	1
homosexual	1
young	1

Interests

The query describes things that interest the user. There were no clear cases of this in L^{WID}; all potential candidates were classified as subjects that the user knows something about, i.e. as skills.

Origin

The query names the nationality or geographic origin of the user. These are among the most common queries of all; the 1st, 3rd, 4th and 6th most frequent query are nationalities. A sub-categorization was made according to their character as nationalities or geographic terms. There were 1% geographic references from outside Denmark and 7% nationalities. The table below contains 25 of the most commonly occurring examples along with 25 unique occurrences, in order to demonstrate the sort of variation that exists in the low end of a search log.

Query	Subclass	Freq.	Query	Subclass	Freq.
spanish	Nationality	2.419	**kineser** 'Chinese (person)'	Nationality	1
french	Nationality	1.411	**koszalin**	Geographic	1
russian	Nationality	1.379	**latinoamerica**	Geographic	1
english	Nationality	1.273	**letisk** '[Latvian]'	Nationality	1
italian	Nationality	987	**Macedonia**	Geographic	1
polish	Nationality	591	**makedonska** (Macedonian) 'Macedonian'	Nationality	1

german	Nationality	529	**mexicano** (Spanish) 'Mexican'	Nationality	1
finnish	Nationality	508	**nordmand** 'Norwegian'	Nationality	1
portuguese	Nationality	451	**Poish** [Polish]	Nationality	1
chinese	Nationality	382	**polih** [Polish]	Nationality	1
japanese	Nationality	269	**rumania**	Geographic	1
spansk 'Spanish'	Nationality	227	**russen** (German?) 'Russians'	Nationality	1
arabic	Nationality	225	**russiak** [Russian]	Nationality	1
dutch	Nationality	224	**slovakisk** 'Slovak'	Nationality	1
china	Geographic	214	**soviet union**	Geographic	1
Russia	Geographic	173	**spnai** [Spain]	Geographic	1
japan	Geographic	169	**swdish** [Swedish]	Nationality	1
turkish	Nationality	163	**turkj** [Turk]?	Nationality	1
English	Nationality	151	**tysker** 'German'	Nationality	1
poland	Geographic	144	**ukrsine** [Ukraine]	Geographic	1
swedish	Nationality	142	**ungary** [Hungary]	Geographic	1
hungarian	Nationality	140	**ungrn** '[Hungary]'	Geographic	1
romanian	Nationality	138	**welsh**	Nationality	1
German	Nationality	137	**zaukraine** (Ukrainian) 'Ukrainian'	Nationality	1
tysk 'German'	Nationality	126	**zurich**	Geographic	1

The language used is included in parantheses whenever it is neither Danish nor English. Incorrectly spelled words are given with their interpretation in brackets.

- **Situation**

The query describes the problematic situation that triggered the search. Examples are *unemployed* and *I'm looking for a job*.

Query	Frequency
unemployed	9
busco trabajo (Spanish) 'I am looking for a job'	1
i'm looking for a job	1
jeg søger en arbejdeskiver i Denmark 'I am looking for an [employer] in [Denmark]'[225]	1
jobseeker	1

Experience

The query describes the user's previous occupations or certificates earned. As job titles were consistently interpreted as desired jobs, experience was mostly restricted to expressions of limited experience such as *no experience*, *unskilled*, and *new graduate*.

Query	Frequency
no experience	21
unskilled	18
new graduate	7
newly graduate	7
without experience	4
recently graduate	3
none experience	1
nyuddannet 'newly graduate'	1
PostGraduate	1
Recent graduate	1
un skilled worker	1

Skills

The class of skills covers 6% abilities (using a tool, teaching a subject, linguistic skills, etc.) and 2% subjects that the user knows something about. Examples are *java*, *forklift*, *photoshop*, *turkish speaking* and *urban planning*.

Query	Subclass	Freq.	Query	Subclass	Freq.
java	Tool	913	**tsm**	Tool	1
sap	Tool	554	**turkish speaking**	Linguistic	1
php	Tool	452	**tysk talende** 'German-speaking'	Linguistic	1
SAP	Tool	360	**udfoldelse af produktet** 'unfolding of the product'	Subject	1

[225] The brackets indicate a misspelled word.

c	Tool	342	**uml**	Tool	1
oracle	Tool	312	**urban palning** [urban planning]	Subject	1
.net	Tool	288	**Urban Planning**	Subject	1
cad	Tool	236	**usability** [usability]	Subject	1
GIS	Tool	193	**use experience** [user experience]	Subject	1
linux	Tool	186	**utility**	Subject	1
autocad	Tool	183	**validaton** [validation]	Subject	1
C	Tool	183	**VFX**	Tool	1
cnc	Tool	174	**vibration turbine**	Subject	1
gis	Tool	174	**video coding**	Subject	1
catia	Tool	150	**VSAT**	Tool	1
network	Tool	146	**wealth manangement** [wealth management]	Subject	1
sql	Tool	128	**web dynpro**	Tool	1
office	Tool	119	**wind mechanical engineering**	Subject	1
mainframe	Tool	118	**windchill**	Subject	1
javascript	Tool	116	**windkraft** (German) 'wind power'	Subject	1
forklift	Tool	115	**WINForms**	Tool	1
photoshop	Tool	114	**wózki widlowe** (Polish) 'forklift'	Tool	1
3d	Tool	92	**wte qa qc**	Tool	1
asp.net	Tool	88	**xml, java, sql**	Tool	1
abap	Tool	87	**zk**	Tool	1

● Schooling

This class was supposed to cover references to the kind of institution were the user received his or her education, and the theoretical school followed. There were no clear instances of this in L^{WID}.

● Field

The field class covers 5% educational titles such as *chemical engineer* and *biologist*, and 4% academic fields, e.g. *biology*. In L^{WID}, these are almost exclusively in English.

Query	Subclass	Freq.	Query	Subclass	Freq.
engineer	Title	1.581	**Social Science**	Academic	1
civil engineer	Title	640	**SOCIAL SCIENCES**	Academic	1
mechanical engineer	Title	600	**Social scientist**	Title	1
chemistry	Academic	399	**Sociolog**	Title	1

biology	Academic	347	**Software build engineer**	Title	1
electrical engineer	Title	309	**Software ENgineer**	Title	1
engineering	Academic	300	**software testengineer**	Title	1
chemical engineer	Title	268	**sociology**	Academic	1
law	Academic	260	**speech logopedi**	Academic	1
biotechnology	Academic	234	**Speech Pathologist**	Title	1
chemical engineering	Academic	225	**STATISTICS**	Academic	1
statistics	Academic	215	**statistics and econometrics**	Academic	1
economics	Academic	196	**stuctural engineer**	Title	1
physics	Academic	193	**support engineer**	Title	1
geology	Academic	187	**tandem systems engineer**	Title	1
molecular biology	Academic	181	**Technical chemistry**	Academic	1
mechanical engineering	Academic	161	**technical health ambiental**	Title	1
process engineer	Title	159	**technical ingeneer**	Title	1
technician	Title	156	**Traffic engineer**	Title	1
psychiatry	Academic	144	**transmission line engineer**	Title	1
psychology	Academic	144	**vaccinology**	Academic	1
microbiology	Academic	132	**vas Engineer**	Title	1
civil engineering	Academic	131	**vibration measurement engineer**	Title	1
structural engineer	Title	125	**water enginneer**	Title	1
anthropology	Academic	116	**wind expert**	Title	1

- **Need**

The need class covers basic descriptions of what is sought, e.g. *job*, *internship*, *student work*, etc. It also covers terms of seniority, as these likewise ensure that the job is one that suits the specific needs of the user rather than merely being more or less desirable.

Query	Subclass	Freq.	Query	Subclass	Freq.
internship	Need type	619	**jobsa** [jobs]	Need type	1
trainee	Seniority	385	**junier** [junior]	Seniority	1
student job	Need type	242	**labopr** [labour]	Need type	1
junior	Seniority	171	**løntilskudjob** 'subsidized job'	Need type	1
job	Need type	87	**msproject**	Need type	1
entry level	Seniority	79	**paid internship**	Need type	1
work	Need type	73	**PHD POSITIONS**	Need type	1
student jobs	Need type	58	**phd vacant position**	Need type	1

all	Need type	55	**post doctoral job**	Need type	1
labour	Need type	51	**postdocrotral**	Need type	1
jobs	Need type	47	**semi-faglært arbejde** 'semi-skilled work'	Need type	1
intership	Need type	38	**Sommerferieafløser** 'summer replacement'	Need type	1
any	Need type	36	**student ass*stant**	Need type	1
student assistant	Need type	34	**student assistance**	Need type	1
Internship	Need type	33	**Student Helper**	Need type	1
internships	Need type	23	**student position**	Need type	1
all jobs	Need type	22	**student weekend**	Need type	1
praktik	Need type	21	**Studentermedhjælper** 'student helper'	Need type	1
hot jobs	Need type	20	**Students Jobs**	Need type	1
labor	Need type	17	**totti lavori** (Italian) 'all jobs'	Need type	1
student helper	Need type	15	**Trainee program**	Need type	1
studentjob	Need type	15	**trainee programme**	Need type	1
entry	Seniority	15	**volunteer jobs**	Need type	1
any job	Need type	14	**whatever**	Need type	1
praktikant 'intern'	Need type	12	**wor** [work]	Need type	1

⬤ Methods

This class encompasses systems, tools and methods that the user wishes to work with (as opposed to those reported as skills), such as *web*, (*job where to drive*) *forklift*, etc. As tools were classed as skills unless there were specific reasons to include them here, this class is very limited, but the following queries were tentatively included, among others.

Query	Subclass	Freq.	Query	Subclass	Freq.
web	System	182	**commercial operation**	Method	1
android	System	60	**communicaton design**	Method	1
iphone	System	42	**conservation and restoration**	Method	1
ASIC	System	17	**constructionengineer ing**	Method	1
asic	System	12	**Coordination**	Method	1
biostatistics	Method	12	**cost estimate**	Method	1
control system	System	11	**crisis managemetn**	Method	1
ios	System	11	**critical care**	Method	1
networks	System	10	**DIALYSIS**	Method	1
mass spectrometry	Method	9	**electroni**	System	1
offset	System	9	**formwork**	System	1
Android	System	8	**ground penetrating radar**	System	1

coldfusion	Method	8	**LANGUAGE TECHNOLOGY**	System	1
Web	System	8	**LEAN**	Method	1
net	System	7	**macintosh**	System	1
3G	System	6	**måleteknikker** ‘measuring techniques’	Method	1
analysis	Method	6	**mamography**	Method	1
virtualization	Method	6	**onlin**	System	1
financial modeling	Method	5	**power electronics research**	Method	1
GPS	System	5	**shiatsu**	Method	1
medical imaging	Method	5	**using Japanese,chinese or english**	System	1
bioanalytical	Method	4	**water hammer**	System	1
chromatography	Method	4	**wcag**	Method	1
clinical trial	Method	4	**with gravemaskine** ‘with excavator’	System	1
market analysis	Method	4	**x509**	Method	1

⬤ Locality

The locality class consists of place names within the confines of the Danish monarchy (incl. Greenland and the Faeroe Islands). Most place names are not assignable to any specific language (this is also true if they are clearly of Danish origin, since few place names have distinct international forms), but a few occur in the log in an English, German, or even Lithuanian form. The fact that Copenhagen is only the third most common city is probably due to its being used in several different forms depending on language (and also because of spelling difficulties).

Query	Freq.	Query	Freq.
aarhus	884	**kolt**	1
aalborg	664	**kopenhagien**	1
copenhagen	646	**Kopenhgagen**	1
odense	562	**koppenhagen**	1
herning	322	**lejre**	1
horsens	285	**lønstrup**	1
esbjerg	251	**mid jutland**	1
Aalborg	216	**Midjytland**	1
Aarhus	191	**midylland**	1
kolding	191	**Næstved**	1
vejle	149	**ondense**	1
Copenhagen	123	**praesto**	1
Kopenhagen	116	**rinkiobing**	1
danmark	109	**rømø-**	1
arhus	105	**skærbæk**	1
århus	96	**skørping**	1

Odense	95	**soendebrorg**	1
denmark	90	**sonder-felding**	1
viborg	89	**sonnenborgh**	1
randers	82	**south denmark**	1
Horsens	67	**strue**	1
fredericia	57	**Them. Midtjylland**	1
Herning	54	**thyboron**	1
sonderborg	54	**tisted**	1
roskilde	51	**veholstebro**	1

⬤ Conditions

The conditions class encompasses desired working hours, type of salary, periodic work etc. It is a group of limited size. Examples are *part time*, *flex*, *seasonal work*, *volunteer* and *freelancer*.

Query	Subclass	Freq.	Query	Subclass	Freq.
part time	Duration	235	**volunteering**	Conditions	3
project	Duration	121	**parttime job**	Duration	2
summer	Duration	114	**project assistance**	Duration	2
seasonal work	Duration	102	**season works**	Duration	2
seasonal	Duration	92	**sesonalwork**	Duration	2
part-time	Duration	76	**afternoon**	Duration	1
summer job	Duration	54	**ÃƒÂ¢Ã¢â€šÂ¬Ã…â€œpart timeÃƒÂ¢Ã¢â€šÂ¬Ã…â€œ**	Duration	1
parttime	Duration	46	**frelancer**	Conditions	1
temporary	Duration	46	**julemåneder** 'christmas months'	Duration	1
season work	Duration	32	**parti time**	Duration	1
seasonal job	Duration	27	**pert-time job**	Duration	1
volunteer	Conditions	26	**sæsonarbejdskraft** 'seasonal labour'	Duration	1
part time job	Duration	20	**seasonla job**	Duration	1
deltid 'part time'	Duration	19	**seazon job**	Duration	1
flex	Duration	17	**semi employment**	Duration	1
seasonalwork	Duration	17	**Sezonas darbi** (Latvian) 'seasonal work'	Duration	1
part-time job	Duration	16	**sezonwork**	Duration	1
part time jobs	Duration	14	**sommerjobs**	Duration	1
season job	Duration	14	**summer season**	Duration	1
seasonal jobs	Duration	14	**summer trainee**	Duration	1
summer jobs	Duration	14	**temp work**	Duration	1
summerjob	Duration	13	**temporary worker**	Duration	1
Project	Duration	11	**two month**	Duration	1
sæsonarbejde 'seasonal work'	Duration	11	**week end**	Duration	1
sesonal	Duration	9	**weekend job**	Duration	1

⬤ Goal

The goal class covers 26% job titles and 5% work task descriptions such as *cleaning*, *pick strawberries* and *drive truck*. Most queries are in English, but some are in other languages.

Query	Subclass	Freq.	Query	Subclass	Freq.
driver	Job title	1.205	**VTS SEA CONTROLER**	Job title	1
cleaning	Task	1.156	**wach disses** [wash dishes]	Task	1
teacher	Job title	887	**washing-up**	Task	1
nurse	Job title	853	**watresing** [waitressing]	Task	1
architect	Job title	831	**weaving**	Task	1
chemist	Job title	567	**Web Design**	Task	1
designer	Job title	527	**WEB DESIGNER**	Job title	1
cleaner	Job title	491	**WEB DEVELOPER**	Job title	1
accountant	Job title	482	**web deweloper**	Job title	1
welder	Job title	482	**web programør**	Job title	1
waiter	Job title	480	**Webdesigner**	Job title	1
cook	Job title	425	**welder.**	Job title	1
lawyer	Job title	407	**welding manager**	Job title	1
dentist	Job title	398	**Welter**	Job title	1
electrician	Job title	392	**werlder**	Job title	1
graduate	Job title	387	**wind farm control**	Job title	1
project manager	Job title	379	**window cleaner**	Job title	1
worker	Job title	368	**windows fitter**	Job title	1
chef	Job title	360	**windows system adminsitrator**	Job title	1
doctor	Job title	334	**wine salesman**	Job title	1
carpenter	Job title	306	**wine somielier**	Job title	1
analyst	Job title	301	**work site manager**	Job title	1
pharmacist	Job title	301	**workew** [worker]	Job title	1
postdoc	Job title	263	**workwr** [worker]	Job title	1
painter	Job title	252	**wraiters** [writers]	Job title	1

⬤ Core

The core class contains products, materials and work focus, such as *child*, *safety*, *pig* and *iron*. Most people use English.

Query	Subclass	Freq.	Query	Subclass	Freq.
water	Material	386	**susi** [sushi]	Material	1
food	Material	341	**suvenior** [souvenir]	Material	1
pig	Material	267	**swimming pool**	Material	1
quality	Focus	257	**Tand** 'tooth'	Material	1
security	Focus	233	**texlite clothing**	Material	1
pigs	Material	232	**thatched roofs**	Material	1
soil	Material	232	**tide**	Material	1
music	Material	226	**tobacco**	Material	1
mink	Material	218	**tool cnc**	Material	1
software	Material	215	**trash**	Material	1
supply chain	Focus	201	**uninterruptible power**	Material	1
cow	Material	156	**vedligeholdelse** 'maintenance'	Focus	1
art	Material	154	**veje** 'roads'	Material	1
electronics	Material	154	**vending machines**	Material	1
sustainability	Focus	133	**Veterinar medicine**	Material	1
car	Material	128	**villa**	Material	1
wind	Material	114	**vod**	Material	1
computer	Material	108	**wind composite**	Material	1
sushi	Material	104	**WINDOWS AND DOORS**	Material	1
safety	Focus	96	**wood furniture**	Material	1
innovation	Focus	88	**wood house**	Material	1
wind turbine	Material	88	**woods**	Material	1
fish	Material	87	**workforce**	Focus	1
textile	Material	85	**x-yachts**	Material	1
strategy	Focus	84	**yogurt**	Material	1

⬤ Workplace

The workplace class consists of 3% company names and 4% workplaces such as *farm*, *offshore*, *hotel* and *museum*. Most people use English terms.

Query	Subclass	Freq.	Query	Subclass	Freq.
farm	Environment	1.298	**top-toy**	Company	1
hotel	Environment	656	**toshiba**	Company	1
warehouse	Environment	259	**trustpilot**	Company	1
restaurant	Environment	258	**Trustpilot ApS**	Company	1
bank	Environment	211	**Trykkeri**	Environment	1
factory	Environment	193	**ujságos** (Hungarian) '[newsagent]'	Environment	1
vestas	Company	166	**újságos** (Hungarian) 'newsagent'	Environment	1

cisco	Company	142	**ukraine Alfa**	Company	1
laboratory	Environment	142	**unidrain**	Company	1
kitchen	Environment	136	**Universities**	Environment	1
maersk	Company	136	**university of aalborg**	Company	1
lab	Environment	132	**ursing home** [nursing home]	Environment	1
NGO	Type	128	**Vejle Business College**	Company	1
bar	Environment	127	**verigy**	Company	1
library	Environment	125	**veterinary office**	Environment	1
pig farm	Environment	125	**vmax**	Company	1
pigs farm	Environment	113	**vw**	Company	1
siemens	Company	112	**warenhous** (German) '[warehouse]'	Environment	1
hotels	Environment	111	**web shop**	Environment	1
farms	Environment	109	**weber**	Company	1
pharmacy	Environment	105	**westas**	Company	1
lego	Company	103	**WIND FARMS**	Environment	1
airport	Environment	97	**wirehouse** [warehouse]	Environment	1
dairy	Environment	94	**yard**	Environment	1
lager 'storage'	Environment	90	**zos** [zoos]	Environment	1

⬤ Sector

The last broad class, sector, covers 8% true sectors and 3% departments and activities that are performed across company types, such as *sales*, *hr*, *management*, etc. Most searches are in English.

Query	Subclass	Freq.	Query	Subclass	Freq.
marketing	Dep.	997	**process industry**	Area	1
finance	Area	600	**Psy**	Area	1
agriculture	Area	588	**public works**	Area	1
it	Area	568	**QA and Testing**	Dep.	1
IT	Area	540	**rchitecture**	Area	1
communication	Area	507	**sales and customer care**	Dep.	1
design	Dep.	475	**sales, procurement, purchasing**	Dep.	1
energy	Area	458	**sales,purchase and marketing**	Dep.	1
sales	Dep.	454	**Sales,purchase and Marketing (69)**	Dep.	1
HR	Dep.	388	**sanitary**	Area	1
legal	Area	386	**seefahrt** (German) 'seafaring'	Area	1
environment	Area	348	**service teknik**	Area	1
environmental	Area	326	**shiping and logistics**	Dep.	1

construction	Area	298	**shipping hseq**	Dep.	1
architecture	Area	255	**shipping/receiving**	Dep.	1
logistics	Dep.	249	**Shoe Industry**	Area	1
accounting	Dep.	240	**Spa and beauty**	Area	1
tourism	Area	230	**Tele**	Area	1
hr	Dep.	206	**telekomunication**	Area	1
administration	Dep.	194	**telephony**	Area	1
media	Area	162	**textile business**	Area	1
telecom	Area	162	**træ industry** 'wood industry'	Area	1
education	Area	159	**transpor**	Area	1
fashion	Area	158	**veterib**	Area	1
transport	Area	156	**Werkzaamheden op de productie van** (Dutch) 'work on the production of'	Area	1

Appendix D1 – Anticipated verbs in L^{TDC}

This appendix lists present and past verbs in queries from L^{TDC} with a frequency above 10. Most of the present forms ending in *-rer* are likely to represent spelling mistakes as they are pronounced identically to the corresponding infinitives in *-re* and this is a common spelling error in Danish. The rest are likely candidates for anticipated text.

Query	Translation	Tense	Frequency
viderestiller	diverts	present	1007
fejlmelder	reports errors	present	298
hjalp	helped	past	276
downloader	downloads	present (/agent?)	172
indtaster	enters (characters)	present	160
aktiverer	activates	present (/infinitive)	89
gav	gave	past	84
fungerer	works	present (/infinitive)	83
vurderer	assesses	present (/infinitive)	65
installerer	installs	present (/infinitive)	58
tilslutter	connects	present	44
opsiger	terminates (subscription)	present	39
klikker	clicks	present	33
frigiver	releases	present	31
abonnerer	subscribes	present (/infinitive)	30
signalerer	signals	present (/infinitive)	27
besvarer	answers	present (/infinitive)	26
aflytter	listen to (voicemail)	present	19
angav	reported	past	18
henter	loads	present	17
deaktiverer	deactivates	present (/infinitive)	16
flytter	moves	present	12
korrigerer	corrects	present (/infinitive)	11

Appendix D2 – Verb use in L^{TDC}

The following table summarizes the use of various verbal forms in L^{TDC}. Only verbs that occur initially in queries with a frequency above 10 are included. Clearly dominant forms are shown with their frequencies in bold. The *Other* category includes tensed forms and agent nouns, but also forms such as *klage* which are ambiguous between a verbal noun and an infinitive. For reasons discussed in section 9.4, such cases are not treated as *Ambiguous* here because verbal nouns and infinitives pattern together in the analyses performed there and thus need not be kept so strictly apart as other cases of formal ambiguity.

Verb	V.Imp	V.Inf	V.nom	Other	Ambiguous	n
oplyse 'inform'	0,26%		**99,74%**			83.758
opsige 'cancel'	4,69%	12,88%	**82,34%**	0,10%		41.032
teste 'test'	4,73%	0,13%	2,27%		**92,87%**[226]	31.172
opsætte 'configure'	8,01%	2,58%	**88,36%**	1,06%		18.988
finde 'find'	**99,71%**	0,29%				13.866
fejlmelde 'report error'	6,80%	8,08%	**82,70%**	2,43%		12.282
flytte 'move'	8,11%	6,94%	**84,82%**	0,13%		9.283
viderestille 'redirect'	4,69%	10,63%	**71,98%**	12,70%		8.739
oprette (user) 'create'	**80,96%**		19,04%			478
oprette (mailaddress)	**74,68%**	20,05%	5,27%			2.788
oprette (homepage)	**65,25%**	21,19%	13,56%			118
oprette (directory/mail /program)	**76,68%**	15,00%	8,31%			2.646
oprette (redirect service/web mail)	**100,00%**					260
oprette ()	17,02%		**62,61%**	20,36%		329
betale 'pay'	**69,33%**		30,67%			5.617
optanke 'prepay'	7,54%	0,41%	**92,05%**			5.345
klage 'complain'	1,92%			**98,08%**		5.165
se 'watch'	2,21%	4,55%	1,97%		**91,27%**[227]	4.261
abonnere 'subscribe'			**97,36%**	2,64%		3.784
afmelde 'cancel (subscription)'	39,21%	12,32%	48,47%			3.466
ændre 'change'		**66,83%**	33,17%			2.674
sende (complaint) 'send'	**71,90%**	28,10%				210
sende (mail)	**81,50%**	18,50%				1.973
sende ()	**100,00%**					19
downloade 'download'		2,71%	6,79%	9,13%	**81,37%**	1.884
omstille 'redirect'	3,25%	0,87%	**95,88%**			1.844
installere 'install'	32,20%	5,65%	**57,20%**	4,94%		1.680

[226] The form *test* can be either a verbal noun or an imperative in Danish.
[227] Robbed of its context, *se* can be either an imperative or an infinitive. Whenever there was a context, the proper interpretation was registered.

måle ’measure’	**73,15%**	8,23%	12,41%	6,21%		1.531
aktivere ’activate’	17,42%	23,80%	**52,48%**	6,30%		1.412
søge (person) ’search’	**100,00%**					804
søge ()	26,00%	38,74%	15,71%	19,55%		573
købe ’buy’		3,59%	18,10%		**78,32%**	1.365
overdrage ’transfer’	1,73%	3,15%	**95,12%**			1.270
reparere ’repair’	4,83%		**95,17%**			1.036
opfylde ’prepay’		1,93%	**98,07%**			1.034
leje ’rent’	**63,76%**		14,37%	21,87%		974
udvide ’expand (mailbox)’	**71,87%**	11,28%	16,86%			949
lukke ’close’	30,52%	18,33%	45,23%	5,93%		911
opdatere ’update’	4,61%	1,54%	**93,85%**			911
indstille ’configure’		2,43%	**97,57%**			905
tilmelde ’sign up’	**79,18%**	3,47%	17,35%			807
hjælpe ’help’			21,90%	34,94%	43,16%	790
slette ’delete’	**68,54%**	25,35%	6,11%			785
spørge ’ask’	**78,69%**		21,31%			643
registrere ’register’	23,52%		**76,48%**			574
tilslutte ’connect (device)’			**91,67%**	8,33%		528
aflytte ’listen (to voicemail)’	**70,53%**	25,86%		3,61%		526
bestille ’order’	**63,10%**	11,09%	25,81%			523
scanne ’scan’	**76,36%**	2,33%		21,32%		516
skrive (complaint) ’write’	**93,35%**	6,65%				361
skrive (mail)	**100,00%**					103
skrive ()	**100,00%**					21
afbestille ’unsubscribe’	22,15%	10,05%	**67,81%**			438
blokere ’block’	39,14%	25,20%	35,66%			373
kryptere ’encrypt’		3,24%	**96,76%**			370
modtage ’receive’	22,95%	38,52%		38,52%		366
åbne ’open’		36,91%	**63,09%**			317
opgradere ’upgrade’	35,74%	8,25%	**56,01%**			291
optage ’record’	19,43%	**60,78%**	4,59%	15,19%		283
hente (content) ’load’	**100,00%**					200
hente ()		**76,71%**		23,29%		73
melde ’report’	**65,17%**	34,83%				267
indtaste ’enter (information)’	6,77%	17,67%		**75,56%**		266
uploade ’upload’		29,23%			**70,77%**	260
godkende ’accept’	**71,43%**		28,57%			245
synkronisere ’synchronize’	5,81%	**60,17%**	34,02%			241
læse ’read’	**53,33%**	16,67%		30,00%		240
vælge ’choose’	**67,84%**	4,85%	27,31%			227
ophæve ’repeal’	**74,65%**	15,49%	9,86%			213

overføre ’transfer’	11,82%	12,32%	**75,86%**			203
begrænse ’limit’		**50,51%**	49,49%			198
surfe ’surf’	39,25%	**43,55%**		17,20%		186
fremkalde ’develop (photos)’	27,43%	9,71%	**62,86%**			175
deaktivere ’deactivate’	**71,68%**		6,94%	21,39%		173
tømme ’empty’	**84,39%**	15,61%				173
fjerne ’remove’	**100,00%**					161
tilknytte ’attach’	**100,00%**					154
geninstallere ’re-install’	23,13%		**76,87%**			134
genåbne ’re-open’		18,80%	**81,20%**			133
konfigurere ’configure’	17,19%	23,44%	**59,38%**			128
påfylde ’prepay’	18,33%		**81,67%**			120
fortryde ’cancel (action)’	11,50%		**88,50%**			113
blive ’become’	**100,00%**					110
klikke ’click’		**67,00%**		33,00%		100
beholde ’keep’	30,11%	**69,89%**				93
dobbeltklikke ’double click’		34,09%			**65,91%**	88
etablere ’establish’	**74,12%**		25,88%			85
indsætte ’insert’	**100,00%**					85
redigere ’edit’	**100,00%**					85
give ’give’				**100,00%**		84
svare ’answer’		22,62%		**51,19%**	26,19%	84
beskytte ’protect’	**61,45%**		38,55%			83
fungere ’function’				**100,00%**		83
sammenligne ’compare’	**100,00%**					80
videresende ’forward’	**100,00%**					77
bytte ’exchange’	**68,42%**	31,58%				76
uddybe ’elaborate’	**100,00%**					75
besvare ’answer’		19,18%	45,21%	35,62%		73
spole ’wind’	**100,00%**					73
sikkerhedskopiere ’back up’		42,42%	**57,58%**			66
vurdere ’estimate’				**100,00%**		65
gendanne ’restore’	**52,46%**			47,54%		61
gøre ’do’	**100,00%**					61
følge ’follow’	**74,58%**	25,42%				59
udskrive ’print’	22,41%		**77,59%**			58
lokke ’lure’		**59,65%**		40,35%		57
oversætte ’translate’	**75,44%**		24,56%			57
framelde ’unsubscribe’	44,64%		**55,36%**			56
misbruge ’abuse’		**54,55%**		45,45%		55
henvende ’contact’	**100,00%**					53
invitere ’invite’	31,37%		**68,63%**			51

nulstille 'reset'	**78,43%**		21,57%		51
nybygge 'build'	30,61%		**69,39%**		49
optimere 'optimize'	37,50%		**62,50%**		48
rette 'correct'		**51,16%**	48,84%		43
afinstallere 'uninstall'	**100,00%**				32
glem 'forget'	**100,00%**				32
frigive 'release'				**100,00%**	31
hacke 'hack'		**58,06%**		41,94%	31
udforske 'explore'	**64,52%**	35,48%			31
signalere 'signal'				100,00%	27
forstyrre 'disturb'	**100,00%**				26
afvente 'await'	**100,00%**				25
løse 'solve'	**100,00%**				25
synes 'think'				**100,00%**	25
tilføje 'add'	**100,00%**				25
prøve 'try'	**100,00%**				20
forlade 'leave'	**100,00%**				19
afgive 'deliver'		**100,00%**			18
angive 'specify'				**100,00%**	18
oploade 'upload'	**100,00%**				18
planlægge 'plan'	**100,00%**				17
afslutte 'terminate'		**100,00%**			16
gemme 'save'	**100,00%**				16
inficere 'infect'		**100,00%**			16
forberede 'prepare'	**100,00%**				14
forstå 'understand'	**100,00%**				14
huske 'remember'	**100,00%**				14
nyde 'enjoy'	**100,00%**				14
slukke 'turn off'	**100,00%**				13
anmelde 'report'	**100,00%**				12
lave 'make'		**100,00%**			12
nedlægge 'discontinue'	**100,00%**				12
udelade 'omit'	**100,00%**				12
korrigere 'correct'				**100,00%**	11
udføre 'perform'	**100,00%**				11
udskifte 'replace'	**100,00%**				11
unzippe 'unzip'	**100,00%**				11
vedhæfte 'attach'	**100,00%**				11

There were few unambiguous tensed forms, but they included the following (surprisingly frequent) pasts.

hjalp	276
gav	84
angav	18

Also included were the following presents.

viderestiller	1007		**klikker**	33
fejlmelder	298		**frigiver**	31
downloader	172		abonnerer	30
indtaster	160		signalerer	27
aktiverer	89		besvarer	26
fungerer	83		**aflytter**	19
vurderer	65		**henter**	17
installerer	58		deaktiverer	16
tilslutter	44		**flytter**	12
opsiger	39		korrigerer	11

In this case, however, most of the forms ending in *–rer* are likely to be misspelled infinitives, as there is no phonetic difference between the two forms in modern Danish. Even so, the material contains many unambiguous presents. To these may be added the passives, which can all be either presents or infinitives.

viderestilles	103		installeres	25
læses	44		misbruges	25
indtastes	41		synes	25
gendannes	29		lokkes	23
læses	28		deaktiveres	21

Finally, the number of unambiguous imperatives is very great, at 42,165 distributed over 98 different verbs, compared to a mere 15,979 unambiguous infinitives distributed over 140 verbs. The following are unambiguous imperatives.

find	13862		godkend	175		omstil	60		udforsk	20
opret	4891		meld	174		henvend	53		forlad	19
betal	3894		fjern	161		byt	52		indtast	18
opsig	1925		ophæv	159		beskyt	51		opload	18
send	1778		tilknyt	154		reparer	50		optimer	18
opsæt	1520		vælg	154		fremkald	48		planlæg	17
afmeld	1359		blokér	146		følg	44		gem	16
søg	1052		tøm	146		oversæt	43		inviter	16
fejlmeld	835		bliv	110		opdater	42		nybyg	15
flyt	753		opgrader	104		nulstil	40		forbered	14
udvid	682		registrer	103		afinstaller	32		husk	14
tilmeld	639		klag	99		gendan	32		nyd	14
lej	621		afbestil	97		glem	32		synkroniser	14
installér	541		se	94		geninstallér	31		fortryd	13
slet	538		deaktiver	85		forstyr	26		sluk	13
spørg	506		indsæt	85		afvent	25		udskriv	13
skriv	461		rediger	85		frameld	25		anmeld	12
viderestil	410		modtag	84		løs	25		nedlæg	12
optank	403		sammenlign	80		tilføj	25		udelad	12
aflyt	371		videresend	77		overfør	24		udfør	11
bestil	330		uddyb	75		konfigurer	22		udskift	11
luk	278		spol	73		overdrag	22		unzip	11
oplys	215		surf	73		påfyld	22		vedhæft	11
hent	200		etabler	63		roter	22			
upload	184		gør	61		prøv	20			

Appendix D3 – Interrogatives in L^{TDC}

The following table illustrates the various uses of interrogative forms in L^{TDC}.

Question type	With "?"	Without "?"	Total	Example query
HVAD 'what'	145	2691	2836	**hvad skal i bruge mit cpr til?** 'what do you need my [social security number] for?'
HVEM 'who'	16	674	690	**hvem har tdc?** 'who uses tdc?'
HVILKEN 'which'	51	627	678	**hvilket abonnement har jeg?** 'which subscription do I have?'
HVOR 'where'	80	102	182	**hvor finder jeg mit kundenummer?** 'where do I find my customer number?' **kabeltv hvor?** 'cable tv where?'
HVOR (degree) 'how'	27	0	27	**hvor hurtigt internet kan jeg få?** 'how fast internet can I get?'
HVORDAN 'how'	120	2292	2416	**hvordan sætter jeg telefonsvareren til?** 'how do I turn voicemail on?' **oplåsning af mobil hvordan** 'unlocking of mobile how'
HVORLEDES 'how'	2	1	3	**hvorledes finder jeg mit kundenummer fastnet på nettet?** 'how do I find my customer number landline on the internet?'
HVORFOR 'why'	14	221	235	**hvorfor ændrer tdc priserne?** 'why does tdc change the prices?' **abonnement spærret hvorfor?** 'subscription locked why?'
HVORNÅR 'when'	15	236	251	**hvornår kommer fibernet til Roskilde?** 'when does fiber network arrive in Roskilde?'
Full questions	165	(not counted)		**hvornår kommer fibernet til Roskilde?** 'when does fiber network arrive in Roskilde?'
Questioned strings	636	(n/a)		**kundenummer?** 'customer number?'

Appendix E1 – Singular use in problem-describing queries

This Appendix contains the full data behind the analysis in chapter 9.

Query	Translation	Frequency
mobil i stykker	'mobile phone broken'	7
fastnettelefon i stykker	'landline phone broken'	3
mobil istykker	'mobile phone [broken]'	2
rrouter i stykker	'[router] broken'	2
telefon gået i stykker	'telephone broken'	2
kabel gravet i stykker	'cable broken (while digging)'	1
iphon er gået i stykker	'[iphone] is broken'	1
i stykker	'broken'	1
telefonen er i stykker	'the telephone is broken'	1
min telefon er i stykker	'my telephone is broken'	1
min email er i stykker	'my e-mail is broken'	1
iphone i stykker	'iphone broken'	1
simkort er gået i stykker	'sim card is broken'	1
sim kort i stykker	'sim card broken'	1
digital modtager i stykker	'digital receiver broken'	1
skærm i stykker	'monitor broken'	1
min iphone er gået i stykker	'my iphone is broken'	1
mobilen er i stykker	'the mobile phone is broken'	1
e-mail i stykker	'e-mail broken'	1
min mobil er gået i stykker	'my mobile phone is broken'	1
telefon i stykker	'telephone broken'	1
display istykker	'display [broken]'	1
fjernbetjening i stykker	'remote control broken'	1
home trio boksen gået i stykker	'the home trio box broken'	1
dispaly i stykker	'[display] broken'	1

min tkf er gået i stykker	‘my [telephone] is broken’	1	
min tdc boks er i stykker	‘my tdc box is broken’	1	
modem i stykker	‘modem broken’	1	
Gået i stykker	‘Broken’	1	
taletid simkort i stykker	‘prepaid time sim card broken’	1	
istykker	‘[broken]’	1	
fjernbetjening til hometrio i stykker	‘remote control for hometrio broken’	1	

Appendix E2 – Locking and unlocking an iPhone

This Appendix contains examples from L^{TDC} relating to the Source problem of locking or unlocking an iphone. The model denominators *3*, *3g*, *3gs* or *4* have been removed from directly after *iphone* in the examples below, and capitalizations have been neutralized. A few unambiguous typographical errors have been removed for clarity. Some of these queries can be analysed as either referring to the problematic Stage 1 of one task or the desired Stage 2 of the opposite task; they are stadially ambiguous as shown in the formal representations.

Query	Verb form	Frequency
fjern simlås fra iphone 'remove sim lock from iphone' unlocking: {N2: fjern$_{v.imp}$ simlås fra iphone}	imp	2
fjern simlås iphone 'remove sim lock iphone' unlocking: {(N2: fjern$_{v.imp}$ simlås) (iphone)}	imp	8
fjern sim-låsen på din iphone 'remove the sim-lock on your iphone' unlocking: {N2: fjern$_{v.imp}$ sim-låsen på din iphone}	imp	1
fjern sim-låsen på iphone 'remove the sim-lock on iphone' unlocking: {N2: fjern$_{v.imp}$ sim-låsen på iphone}	imp	1
glemt låsekode iphone 'forgotten lock code iphone' unlocking: {(S1: glemt låsekode) (iphone)}	perf	1
hvordan får jeg låst min iPhone op 'how do I unlock my iPhone' unlocking: {N1: hvordan får$_{v.pres}$ jeg låst_op min iphone}	pres	1
hvordan låser man en iphone op 'how do you unlock an iphone' unlocking: {N1: hvordan låser_op$_{v.pres}$ man en iphone}	pres	1
iphone fra tdc simlåst 'iphone from tdc simlocked' unlocking: {S1: iphone fra tdc simlåst} locking: {S2: iphone fra tdc simlåst}	perf	1
iphone ikke simlåst 'iphone not simlocked' locking: {S1: iphone fra tdc ikke simlåst} unlocking: {S2: iphone fra tdc ikke simlåst}	adj	1
iphone lås op 'iphone unlock' unlocking: {(iphone) (N2: lås_op$_{v.imp}$)}	imp	15

iphone låse ‘iphone lock’ locking: {(iphone) (R: låse$_{v.inf}$)}	inf	1
iphone låse op ‘iphone unlock’ unlocking: {(iphone) (R: låse_op$_{v.inf}$)}	inf	2
iphone låst ‘iphone locked’ unlocking: {(S1: iphone låst)} locking: {(S2: iphone låst)}	perf	7
iphone låst pinkode ‘iphone locked pin code’ unlocking: {(S1: iphone låst)} {pinkode}	perf	2
iphone låst sim ‘iphone locked sim’ unlocking: {(S1: iphone låst)} {sim}	perf	2
iphone oplås ’iphone unlock’ unlocking: {(iphone) (N2: oplås$_{v.imp}$)}	imp	4
iphone oplåsning ’iphone unlocking’ unlocking: {(iphone) (R: oplåsning$_{v.nom}$)}	nom	6
iphone siger simlåst ’iphone says simlocked’ unlocking: {(S1: iphone siger simlåst)}	pres	1
iphone sim låse ’iphone sim lock’ locking: {(iphone) (R: sim låse$_{v.inf}$)	inf	1
iphone sim låst ’iphone sim locked’ unlocking: {(iphone) (S1: sim låst)} locking: {(iphone) (S2: sim låst)}	res	7
iphone simlåse ‘iphone simlock’ locking: {(iphone) (R: simlåse$_{v.inf}$)}	inf	1
iphone simlåst ‘iphone simlocked’ unlocking: {(iphone) (S1: simlåst)} locking: {(iphone) (S2: simlåst)}	res	15
iphone låst med pin ’iphone locked with PIN’ unlocking: {(iphone) (S1: låst med pin)} locking: {(iphone) (S2: låst med pin)}	res	1
iphone uden sim lås ’iphone without sim lock’ unlocking: {S2: iphone uden simlås} locking: {S1: iphone uden simlås}	adj	2

iphone lås op 'iphone unlock' unlocking: {(iphone) (N2: lås_op$_{v.imp}$)}	imp	2
iphone låse op 'iphone unlock' unlocking: {(iphone) (R: låse_op$_{v.inf}$)}	inf	1
iphone låsop 'iphone [unlock]' unlocking: {(iphone) (N2: låsop$_{v.imp}$)}	imp	1
iphone op låsning 'iphone un locking' unlocking: {(iphone) (R: op_låsning$_{v.nom}$)}	nom	1
iphone ulåst 'iphone unlocked' unlocking: {S2: iphone ulåst} locking: {S1: iphone ulåst}	adj	1
lås din iphone 'lock your iphone' locking: {R: lås$_{v.imp}$ din iphone}	imp	1
lås din iphone op 'unlock your iphone' unlocking: {R: lås_op$_{v.imp}$ din iphone}	imp	2
lås iphone op 'unlock iphone' unlocking: {N2: lås_op$_{v.imp}$ iphone}	imp	43
lås min iphone op 'unlock my iphone' unlocking: {N2: lås_op$_{v.imp}$ min iphone}	imp	1
lås op af iphone 'unlock of iphone' unlocking: {(N2: lås_op$_{v.imp}$) (af iphone)}	imp	1
lås op iphone 'unlock iphone' unlocking: {(N2: lås_op$_{v.imp}$) (iphone)}	imp	11
lås op sim iphone 'unlock sim iphone' unlocking: {(N2: lås_op$_{v.imp}$ sim) (iphone)}	imp	1
lås sim op iphone 'unlock sim iphone' unlocking: {(N2: lås_op$_{v.imp}$ sim) (iphone)}	imp	1
låse iphone 'lock iphone' locking: {R: låse$_{v.inf}$ iphone}	inf	4
låse iphone op 'unlock iphone' unlocking: {R: låse_op$_{v.inf}$ iphone}	inf	6

låser tdc iphone 'locks tdc iphone' locking: {R: låser$_{v.pres}$ tdc iphone}	pres	1
låsop iphone 'unlock iphone' unlocking: {(N2: låsop$_{v.imp}$) (iphone)}	imp	1
låst iphone 'locked iphone' unlocking: {S1: låst iphone} locking: {S2: låst iphone}	perf	6
låst sim på iphone 'locked sim iphone' unlocking: {S1: låst sim på iphone} locking: {S2: låst sim på iphone}	perf	1
låst simkort i iphone 'locked sim card in iphone' unlocking: {S1: låst simkort i iphone} locking: {S2: låst simkort i iphone}	perf	1
låst simkort iphone 'locked sim card iphone' unlocking: {S1: låst simkort iphone} locking: {S2: låst simkort iphone}	perf	1
min iphone er låst op hvad gør jeg 'my iphone is unlocked what do I do' locking: {N1: min iphone er$_{v.pres}$ låst op hvad gør$_{v.pres}$ jeg}[228]	perf	1
ny iphone låst pinkode 'new iphone locked PIN code' unlocking: {S1: ny iphone låst} {pinkode}	perf	1
oben sim lås iphone '[open] sim lock iphone' unlocking: {(N2: oben$_{v.imp}$ sim lås) (iphone)}	imp	1
op låsning iphone 'un locking iphone' unlocking: {(R: op_låsning$_{v.nom}$) (iphone)}	nom	2
opertørlåst iphone '[operator-locked] iphone' unlocking: {S1: opertørlåst iphone} locking: {S2: opertørlåst iphone}	perf	1
oplås iphone 'unlock iphone' unlocking: {N2: oplås$_{v.imp}$ iphone}	imp	19
oplås sim kort til iphone 'unlock sim card for iphone' unlocking: {N2: oplås$_{v.imp}$ sim kort til iphone}	imp	1

[228] It seems more likely that what this user actually meant was *min iphone er låst hvad gør jeg* 'my iphone is locked, what to do'. However, the analysis shown here is true to the query as issued.

oplåse iphone ’unlock iphone’ unlocking: {R: oplåse$_{v.inf}$ iphone}	inf	1
oplåsning af iphone ’unlocking of iphone’ unlocking: {R: oplåsning$_{v.nom}$ af iphone}	nom	23
oplåsning af en iphone ’unlocking of an iphone’ unlocking: {R: oplåsning$_{v.nom}$ af en iphone}	nom	1
oplåsning iphone ’unlocking iphone’ unlocking: {(R: oplåsning$_{v.nom}$) (iphone)}	nom	7
sim låst af iphone ’sim locked by iphone’ unlocking: {S1: sim låst af iphone} locking: {S2: sim låst af iphone}	perf	1
sim låst iphone ’sim locked iphone’ unlocking: {S1: sim låst iphone} locking: {S2: sim låst iphone}	perf	8
simlåst iphone ’sim-locked iphone’ unlocking: {S1: simlåst iphone} locking: {S2: simlåst iphone}	perf	19
skal have låst min iphone op ’needs to get my iphone unlocked’ unlocking: {S1: skal$_{v.pres}$ have låst min iphone op}	pres	1
åbne en simlåst iphone ’open a sim-locked iphone’ unlocking: {R: åbne$_{v.inf}$ en simlåst iphone}	inf	1

Appendix E3 – Immediate and delayed action

This Appendix contains data relating to the analysis of the use of imperatives in L^{TDC}. The verbs are numbered in order of falling percentages of imperatives and growing proportions of infinitives.

No.	Verb (sense)	Typical context phrases	Total number	With context	Devoid of context
1	oprette (service)	viderestilling, webmail	260	260	0
2	søge (nummer)	mail, tlfnr, person…	804	804	0
3	skrive (mail)	mail	103	103	0
4	hente (indhold)	mms, musik, ringetoner, sikkerhedspakke	200	200	0
5	fjerne	simlås, viderestilling, voicemail	161	161	0
6	tilknytte	mail	154	113	41
7	blive	kontaktet, kunde	110	0	110
8	indsætte	taletid, penge	85	85	0
9	redigere	indstillinger	85	64	21
10	sammenligne	abonnementer, mobil	80	29	51
11	videresende	mail	77	38	39
12	uddybe	-	75	0	75
13	spole	-	73	0	73
14	gøre	til startside	61	0	61
15	henvende	-	53	0	53
16	finde	adresse, forhandler, kundenummer…	13866	13731	135
17	skrive (klage)	klage, til os…	382	361	0
18	tømme	mailboks	173	173	0
19	sende (mail)	mail, sms, mms	1973	1973	0
20	oprette (bruger)	bruger, netbasis, fastnet, telefon…	478	478	0
21	tilmelde	abonnement, betalingsservice…	807	690	117
22	spørge	før du graver, til faktura…	643	47	596
23	nulstille	sikkerhedspakke	51	11	40
24	oprette (mail)	mappe, mail, mms, program…	2646	2646	0
25	scanne	computer, for virus	516	98	418
26	oversætte	-	57	0	57
27	oprette (mailadresse)	mailadresse, freemail…	2788	2788	0
28	ophæve	spærring, saldomaks	213	213	0
29	følge	forbrug, reparation	59	59	0
30	etablere	-	85	0	85
31	måle	bredbånd, hastighed	1531	1256	275
32	sende (klage)	klage	210	210	0
33	udvide	mailboks	949	751	198
34	deaktivere	mobilsvar, voicemail	173	125	48
35	godkende	-	245	0	245
36	aflytte	besked, telefonsvarer	526	446	80
37	betale	regning, faktura…	5617	3860	1757

38	slette	mail, konto, voicemail…	785	729	56
39	bytte	til nyt	76	0	76
40	vælge	telefonnummer	227	227	0
41	oprette (hjemmeside)	hjemmeside	118	118	0
42	melde	flytning	267	267	0
43	leje	film, xbox	974	933	41
44	bestille	kode, simkort, vækning…	523	422	101
45	beskytte	-	83	0	83
46	læse	mail, sms	240	168	72
47	gendanne	slettet mail	61	11	50
48	framelde	-	56	0	56
49	surfe	-	186	0	186
50	afmelde	abonnement, nyhedsbrev, voicemail…	3466	1425	2041
51	blokere	afsender, nummer, opkald	373	154	219
52	opgradere	hastighed, bredbånd	291	92	199
53	installere	sikkerhedspakke, homedisk, bredbånd…	1680	726	954
54	invitere	-	51	0	51
55	lukke	abonnement, mobil, simkort…	911	677	234
56	beholde	nummer, mailadresse	93	93	0
57	fremkalde	billeder	175	84	91
58	søge ()	-	573	68	505
59	registrere	homedisk, nummer	574	230	344
60	geninstallere	sikkerhedspakke	134	18	116
61	modtage	mms, sms i udlandet, mail	366	198	168
62	udskrive	sms	58	13	45
63	afbestille	abonnement, fastnet	438	78	360
64	optage	tv	283	31	252
65	påfylde	taletid, mobil	120	34	86
66	aktivere	simkort, voicemail	1412	745	667
67	konfigurere	mail	128	22	106
68	oprette ()	-	329	0	329
69	overføre	nummer, abonnement	203	125	78
70	fortryde	-	113	0	113
71	flytte	abonnement, adresse, nummer…	9283	2510	6773
72	opsætte	internet, mail, outlook…	18988	12628	6360
73	optanke	mobil, taletid	5345	970	4375
74	fejlmelde	bredbånd, telefon…	12282	799	11483
75	indtaste	-	266	0	266
76	synkronisere	-	241	0	241
77	reparere	mobil, mail…	1036	131	905
78	viderestille	fastnet, mobil, telefon…	8739	1045	7694
79	opsige	abonnement, mobil, sikkerhedspakke…	41032	12995	28037
80	opdatere	mobil, sikkerhedspakke	911	77	834
81	omstille	fastnet, telefon, opkald	1844	357	1487

82	klage	formular, mail, mobil, regning…	5165	146	5019
83	overdrage	abonnement, telefon…	1270	334	936
84	oplyse	118, udland, online…	83758	1218	82540
85	abonnere	fastnet, mobil, simply…	3784	173	3611
86	tilslutte	-	528	0	528
87	give	-	84	0	84
88	fungerer	-	83	0	83
89	vurdere	-	65	0	65
90	opfylde	mobiltid, taletid	1034	277	757
91	indstille	voicemail, mail	905	30	875
92	kryptere	-	370	0	370
93	genåbne	mobil, telefon	133	44	89
94	besvare	-	73	0	73
95	åbne	mms, mobil, port	317	287	30
96	sikkerhedskopiere	-	66	0	66
97	begrænse	-	198	0	198
98	misbruge	-	55	0	55
99	lokke	-	57	0	57
100	ændre	mail, password, pinkode, betaler…	2674	2608	66
101	klikke	-	100	0	100
102	hente ()	-	73	0	73

All verbs irrespective of context

Verbs with ambiguous forms were excluded for clarity.

No.	Verb (sense)	Imperatives	Infinitives	Verb nouns	Other
1	oprette (service)	100,00%	0,00%	0,00%	0,00%
2	søge (nummer)	100,00%	0,00%	0,00%	0,00%
3	skrive (mail)	100,00%	0,00%	0,00%	0,00%
4	hente (indhold)	100,00%	0,00%	0,00%	0,00%
5	fjerne	100,00%	0,00%	0,00%	0,00%
6	tilknytte	100,00%	0,00%	0,00%	0,00%
7	blive	100,00%	0,00%	0,00%	0,00%
8	indsætte	100,00%	0,00%	0,00%	0,00%
9	redigere	100,00%	0,00%	0,00%	0,00%
10	sammenligne	100,00%	0,00%	0,00%	0,00%
11	videresende	100,00%	0,00%	0,00%	0,00%
12	uddybe	100,00%	0,00%	0,00%	0,00%
13	spole	100,00%	0,00%	0,00%	0,00%
14	gøre	100,00%	0,00%	0,00%	0,00%
15	henvende	100,00%	0,00%	0,00%	0,00%
16	finde	99,71%	0,29%	0,00%	0,00%
17	skrive (klage)	93,72%	6,28%	0,00%	0,00%
18	tømme	84,39%	15,61%	0,00%	0,00%
19	sende (mail)	81,50%	18,50%	0,00%	0,00%

20	oprette (bruger)	80,96%	0,00%	19,04%	0,00%
21	tilmelde	79,18%	3,47%	17,35%	0,00%
22	spørge	78,69%	0,00%	21,31%	0,00%
23	nulstille	78,43%	0,00%	21,57%	0,00%
24	oprette (mail)	76,68%	15,00%	8,31%	0,00%
25	scanne	76,36%	2,33%	0,00%	21,32%
26	oversætte	75,44%	0,00%	24,56%	0,00%
27	oprette (mailadresse)	74,68%	20,05%	5,27%	0,00%
28	ophæve	74,65%	15,49%	9,86%	0,00%
29	følge	74,58%	25,42%	0,00%	0,00%
30	etablere	74,12%	0,00%	25,88%	0,00%
31	måle	73,15%	8,23%	12,41%	6,21%
32	sende (klage)	71,90%	28,10%	0,00%	0,00%
33	udvide	71,87%	11,28%	16,86%	0,00%
34	deaktivere	71,68%	0,00%	6,94%	21,39%
35	godkende	71,43%	0,00%	28,57%	0,00%
36	aflytte	70,53%	25,86%	0,00%	3,61%
37	betale	69,33%	0,00%	30,67%	0,00%
38	slette	68,54%	25,35%	6,11%	0,00%
39	bytte	68,42%	31,58%	0,00%	0,00%
40	vælge	67,84%	4,85%	27,31%	0,00%
41	oprette (hjemmeside)	65,25%	21,19%	13,56%	0,00%
42	melde	65,17%	34,83%	0,00%	0,00%
43	leje	63,76%	0,00%	14,37%	21,87%
44	bestille	63,10%	11,09%	25,81%	0,00%
45	beskytte	61,45%	0,00%	38,55%	0,00%
46	læse	53,33%	16,67%	0,00%	30,00%
47	gendanne	52,46%	0,00%	0,00%	47,54%
48	framelde	44,64%	0,00%	55,36%	0,00%
49	surfe	39,25%	43,55%	0,00%	17,20%
50	afmelde	39,21%	12,32%	48,47%	0,00%
51	blokere	39,14%	25,20%	35,66%	0,00%
52	opgradere	35,74%	8,25%	56,01%	0,00%
53	installere	32,20%	5,65%	57,20%	4,94%
54	invitere	31,37%	0,00%	68,63%	0,00%
55	lukke	30,52%	18,33%	45,23%	5,93%
56	beholde	30,11%	69,89%	0,00%	0,00%
57	fremkalde	27,43%	9,71%	62,86%	0,00%
58	søge ()	26,00%	38,74%	15,71%	19,55%
59	registrere	23,52%	0,00%	76,48%	0,00%
60	geninstallere	23,13%	0,00%	76,87%	0,00%
61	modtage	22,95%	38,52%	0,00%	38,52%
62	udskrive	22,41%	0,00%	77,59%	0,00%
63	afbestille	22,15%	10,05%	67,81%	0,00%
64	optage	19,43%	60,78%	4,59%	15,19%
65	påfylde	18,33%	0,00%	81,67%	0,00%

66	aktivere	17,42%	23,80%	52,48%	6,30%
67	konfigurere	17,19%	23,44%	59,38%	0,00%
68	oprette ()	17,02%	0,00%	62,61%	20,36%
69	overføre	11,82%	12,32%	75,86%	0,00%
70	fortryde	11,50%	0,00%	88,50%	0,00%
71	flytte	8,11%	6,94%	84,82%	0,13%
72	opsætte	8,01%	2,58%	88,36%	1,06%
73	optanke	7,54%	0,41%	92,05%	0,00%
74	fejlmelde	6,80%	8,08%	82,70%	2,43%
75	indtaste	6,77%	17,67%	0,00%	75,56%
76	synkronisere	5,81%	60,17%	34,02%	0,00%
77	reparere	4,83%	0,00%	95,17%	0,00%
78	viderestille	4,69%	10,63%	71,98%	12,70%
79	opsige	4,69%	12,88%	82,34%	0,10%
80	opdatere	4,61%	1,54%	93,85%	0,00%
81	omstille	3,25%	0,87%	95,88%	0,00%
82	klage	1,92%	0,00%	0,00%	98,08%
83	overdrage	1,73%	3,15%	95,12%	0,00%
84	oplyse	0,26%	0,00%	99,74%	0,00%
85	abonnere	0,00%	0,00%	97,36%	2,64%
86	tilslutte	0,00%	0,00%	91,67%	8,33%
87	give	0,00%	0,00%	0,00%	100,00%
88	fungerer	0,00%	0,00%	0,00%	100,00%
89	vurdere	0,00%	0,00%	0,00%	100,00%
90	opfylde	0,00%	1,93%	98,07%	0,00%
91	indstille	0,00%	2,43%	97,57%	0,00%
92	kryptere	0,00%	3,24%	96,76%	0,00%
93	genåbne	0,00%	18,80%	81,20%	0,00%
94	besvare	0,00%	19,18%	45,21%	35,62%
95	åbne	0,00%	36,91%	63,09%	0,00%
96	sikkerhedskopiere	0,00%	42,42%	57,58%	0,00%
97	begrænse	0,00%	50,51%	49,49%	0,00%
98	misbruge	0,00%	54,55%	0,00%	45,45%
99	lokke	0,00%	59,65%	0,00%	40,35%
100	ændre	0,00%	66,83%	33,17%	0,00%
101	klikke	0,00%	67,00%	0,00%	33,00%
102	hente ()	0,00%	76,71%	0,00%	23,29%

Verbs with context

No.	Verb (sense)	Imperatives	Infinitives	Verb nouns	Other
1	oprette (service)	100,00%	0,00%	0,00%	0,00%
2	søge (nummer)	100,00%	0,00%	0,00%	0,00%
58	søge ()	100,00%	0,00%	0,00%	0,00%
25	scanne	100,00%	0,00%	0,00%	0,00%
3	skrive (mail)	100,00%	0,00%	0,00%	0,00%
4	hente (indhold)	100,00%	0,00%	0,00%	0,00%
5	fjerne	100,00%	0,00%	0,00%	0,00%
6	tilknytte	100,00%	0,00%	0,00%	0,00%
67	konfigurere	100,00%	0,00%	0,00%	0,00%
8	indsætte	100,00%	0,00%	0,00%	0,00%
9	redigere	100,00%	0,00%	0,00%	0,00%
10	sammenligne	100,00%	0,00%	0,00%	0,00%
11	videresende	100,00%	0,00%	0,00%	0,00%
47	gendanne	100,00%	0,00%	0,00%	0,00%
23	nulstille	100,00%	0,00%	0,00%	0,00%
16	finde	99,71%	0,29%	0,00%	0,00%
17	skrive (klage)	93,35%	6,65%	0,00%	0,00%
34	deaktivere	90,40%	0,00%	9,60%	0,00%
37	betale	90,28%	0,00%	9,72%	0,00%
18	tømme	84,39%	15,61%	0,00%	0,00%
21	tilmelde	84,35%	4,06%	11,59%	0,00%
31	måle	83,28%	9,16%	7,56%	0,00%
19	sende (mail)	81,50%	18,50%	0,00%	0,00%
20	oprette (bruger)	80,96%	0,00%	19,04%	0,00%
36	aflytte	77,58%	22,42%	0,00%	0,00%
24	oprette (mail)	76,68%	15,00%	8,31%	0,00%
46	læse	76,19%	23,81%	0,00%	0,00%
44	bestille	74,88%	10,90%	14,22%	0,00%
27	oprette (mailadresse)	74,68%	20,05%	5,27%	0,00%
28	ophæve	74,65%	15,49%	9,86%	0,00%
29	følge	74,58%	25,42%	0,00%	0,00%
32	sende (klage)	71,90%	28,10%	0,00%	0,00%
33	udvide	68,31%	14,25%	17,44%	0,00%
40	vælge	67,84%	4,85%	27,31%	0,00%
38	slette	67,63%	25,79%	6,58%	0,00%
41	oprette (hjemmeside)	65,25%	21,19%	13,56%	0,00%
42	melde	65,17%	34,83%	0,00%	0,00%
65	påfylde	64,71%	0,00%	35,29%	0,00%
43	leje	64,20%	0,00%	15,01%	20,79%
50	afmelde	48,70%	15,16%	36,14%	0,00%
51	blokere	46,10%	14,29%	39,61%	0,00%
59	registrere	44,78%	0,00%	55,22%	0,00%
61	modtage	42,42%	57,58%	0,00%	0,00%

53	installere	40,22%	11,43%	48,35%	0,00%
55	lukke	37,52%	20,09%	42,39%	0,00%
64	optage	35,48%	64,52%	0,00%	0,00%
57	fremkalde	34,52%	20,24%	45,24%	0,00%
56	beholde	30,11%	69,89%	0,00%	0,00%
52	opgradere	28,26%	13,04%	58,70%	0,00%
77	reparere	26,72%	0,00%	73,28%	0,00%
66	aktivere	26,31%	12,21%	61,48%	0,00%
71	flytte	25,26%	12,35%	62,39%	0,00%
69	overføre	19,20%	20,00%	60,80%	0,00%
74	fejlmelde	15,27%	2,88%	81,85%	0,00%
73	optanke	14,02%	0,00%	85,98%	0,00%
78	viderestille	9,09%	4,69%	86,22%	0,00%
81	omstille	8,96%	0,00%	91,04%	0,00%
79	opsige	7,95%	26,39%	65,66%	0,00%
72	opsætte	5,40%	0,23%	94,37%	0,00%
84	oplyse	0,00%	0,00%	100,00%	0,00%
82	klage	0,00%	0,00%	0,00%	100,00%
85	abonnere	0,00%	0,00%	100,00%	0,00%
80	opdatere	0,00%	0,00%	100,00%	0,00%
91	indstille	0,00%	0,00%	100,00%	0,00%
22	spørge	0,00%	0,00%	100,00%	0,00%
60	geninstallere	0,00%	0,00%	100,00%	0,00%
62	udskrive	0,00%	0,00%	100,00%	0,00%
83	overdrage	0,00%	4,49%	95,51%	0,00%
90	opfylde	0,00%	7,22%	92,78%	0,00%
63	afbestille	0,00%	15,38%	84,62%	0,00%
93	genåbne	0,00%	25,00%	75,00%	0,00%
95	åbne	0,00%	40,77%	59,23%	0,00%
100	ændre	0,00%	67,71%	32,29%	0,00%

Verbs devoid of context

No.	Verb (sense)	Imperatives	Infinitives	Verb nouns	Other
6	tilknytte	100,00%	0,00%	0,00%	0,00%
7	blive	100,00%	0,00%	0,00%	0,00%
9	redigere	100,00%	0,00%	0,00%	0,00%
10	sammenligne	100,00%	0,00%	0,00%	0,00%
11	videresende	100,00%	0,00%	0,00%	0,00%
12	uddybe	100,00%	0,00%	0,00%	0,00%
13	spole	100,00%	0,00%	0,00%	0,00%
14	gøre	100,00%	0,00%	0,00%	0,00%
15	henvende	100,00%	0,00%	0,00%	0,00%
16	finde	100,00%	0,00%	0,00%	0,00%
33	udvide	85,35%	0,00%	14,65%	0,00%
22	spørge	84,90%	0,00%	15,10%	0,00%
38	slette	80,36%	19,64%	0,00%	0,00%
26	oversætte	75,44%	0,00%	24,56%	0,00%
30	etablere	74,12%	0,00%	25,88%	0,00%
23	nulstille	72,50%	0,00%	27,50%	0,00%
35	godkende	71,43%	0,00%	28,57%	0,00%
25	scanne	70,81%	2,87%	0,00%	26,32%
39	bytte	68,42%	31,58%	0,00%	0,00%
45	beskytte	61,45%	0,00%	38,55%	0,00%
43	leje	53,66%	0,00%	0,00%	46,34%
21	tilmelde	48,72%	0,00%	51,28%	0,00%
48	framelde	44,64%	0,00%	55,36%	0,00%
47	gendanne	42,00%	0,00%	0,00%	58,00%
49	surfe	39,25%	43,55%	0,00%	17,20%
52	opgradere	39,20%	6,03%	54,77%	0,00%
51	blokere	34,25%	32,88%	32,88%	0,00%
50	afmelde	32,58%	10,34%	57,08%	0,00%
54	invitere	31,37%	0,00%	68,63%	0,00%
36	aflytte	31,25%	45,00%	0,00%	23,75%
62	udskrive	28,89%	0,00%	71,11%	0,00%
63	afbestille	26,94%	8,89%	64,17%	0,00%
31	måle	26,91%	4,00%	34,55%	34,55%
60	geninstallere	26,72%	0,00%	73,28%	0,00%
53	installere	26,10%	1,26%	63,94%	8,70%
37	betale	23,28%	0,00%	76,72%	0,00%
34	deaktivere	22,92%	0,00%	0,00%	77,08%
57	fremkalde	20,88%	0,00%	79,12%	0,00%
64	optage	17,46%	60,32%	5,16%	17,06%
68	oprette ()	17,02%	0,00%	62,61%	20,36%
58	søge ()	16,04%	43,96%	17,82%	22,18%
44	bestille	13,86%	11,88%	74,26%	0,00%
72	opsætte	13,18%	7,25%	76,42%	3,16%

70	fortryde	11,50%	0,00%	88,50%	0,00%
55	lukke	10,26%	13,25%	53,42%	23,08%
59	registrere	9,30%	0,00%	90,70%	0,00%
66	aktivere	7,50%	36,73%	42,43%	13,34%
75	indtaste	6,77%	17,67%	0,00%	75,56%
74	fejlmelde	6,21%	8,44%	82,76%	2,60%
73	optanke	6,10%	0,50%	93,39%	0,00%
76	synkronisere	5,81%	60,17%	34,02%	0,00%
80	opdatere	5,04%	1,68%	93,29%	0,00%
78	viderestille	4,09%	11,44%	70,04%	14,43%
79	opsige	3,18%	6,61%	90,07%	0,14%
83	overdrage	2,35%	2,67%	94,98%	0,00%
82	klage	1,97%	0,00%	0,00%	98,03%
81	omstille	1,88%	1,08%	97,04%	0,00%
71	flytte	1,76%	4,93%	93,13%	0,18%
77	reparere	1,66%	0,00%	98,34%	0,00%
84	oplyse	0,26%	0,00%	99,74%	0,00%
46	læse	0,00%	0,00%	0,00%	100,00%
65	påfylde	0,00%	0,00%	100,00%	0,00%
69	overføre	0,00%	0,00%	100,00%	0,00%
85	abonnere	0,00%	0,00%	97,23%	2,77%
86	tilslutte	0,00%	0,00%	91,67%	8,33%
87	give	0,00%	0,00%	0,00%	100,00%
88	fungerer	0,00%	0,00%	0,00%	100,00%
89	vurdere	0,00%	0,00%	0,00%	100,00%
90	opfylde	0,00%	0,00%	100,00%	0,00%
91	indstille	0,00%	2,51%	97,49%	0,00%
92	kryptere	0,00%	3,24%	96,76%	0,00%
61	modtage	0,00%	16,07%	0,00%	83,93%
67	konfigurere	0,00%	28,30%	71,70%	0,00%
6	tilknytte	100,00%	0,00%	0,00%	0,00%
7	blive	100,00%	0,00%	0,00%	0,00%
9	redigere	100,00%	0,00%	0,00%	0,00%
10	sammenligne	100,00%	0,00%	0,00%	0,00%
11	videresende	100,00%	0,00%	0,00%	0,00%
12	uddybe	100,00%	0,00%	0,00%	0,00%
13	spole	100,00%	0,00%	0,00%	0,00%
14	gøre	100,00%	0,00%	0,00%	0,00%
15	henvende	100,00%	0,00%	0,00%	0,00%
16	finde	100,00%	0,00%	0,00%	0,00%

References

Aime, X. (2011). *Gradients de prototypicalité, mesures de similarité et de proximité sémantique: une contribution à l'Ingénierie des Ontologies (Doct. Thesis).* Université de Nantes.

Albakour, M.-D., Kruschwitz, U., Nanas, N., Kim, Y., Song, D., Fasli, M., & De Roeck, A. (2011). AutoEval: An evaluation methodology for evaluating query suggestions using query logs. *ECIR* (pp. 605-610). Dublin, Ireland: Springer-Verlag.

Alfort, E. (2007). *Fundamentalt og deiktisk emne.* (University of Copenhagen: unpubl.).

Alfort, E. (2009). *Subjective and objective analyses of referential phenomena in studies of communication and texts (M.A. Thesis).* Copenhagen: University of Copenhagen.

Alfort, E. (2010). Perspective and scale in classification - The classificatory dimension. *TKE*, (pp. 413-433). Dublin, Ireland.

Allan, J., Aslam, J., Belkin, N., & al. (2003). Challenges in information retrieval and language modelling - Report of a workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002. *ACM SIGIR Forum, 37*(1), 31-47.

Anick, P. (2003). Using terminological feedback for web search refinement: A log based study. *SIGIR* (pp. 88-95). Toronto, Canada: ACM.

Arampatzis, A. T., van der Weide, T. P., van Bommel, P., & Koster, C. H. (2000). *Linguistically-motivated Information Retrieval (Technical Report CSI-R9918, September 1999).* Nijmegen, Netherlands: Dept. of Information Systems, Faculty of Mathematics and Computing Science, University of Nijmegen.

Arampatzis, A., Zagoris, K., & Chatzichristofis, S. A. (2011). Dynamic Two-Stage Image Retrieval from Large Multimodal Databases. *ECIR* (pp. 326-337). Dublin, Ireland: Springer Verlag.

Azzopardi, L. (2009). Query side evaluation - An empirical analysis of effectiveness and effort. *SIGIR* (pp. 556-563). Boston, MA: ACM.

Balog, K., Weerkamp, W., & de Rijke, M. (2008). A few examples go a long way - Constructing query models from elaborate query formulations. *SIGIR* (pp. 371-378). Singapore: ACM.

Barnhart, R. K. (1988). *Chambers Dictionary of Etymology.* Edinburgh/New York: Chambers.

Barr, C., Jones, R., & Regelson, M. (2008). The Linguistic Structure of English Web-Search Queries. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 1021-1030). Honolulu: ACL.

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science, 45*(3), 149-159.

Barry, C. L., & Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management, 34*(2/3), 219-236.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition, 11*(3), 211-227.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*, 407-424.

Bateson, G. (1973). *Steps to an ecology of mind.* St. Albans, Australia: Paladin.

Behrens, L. (2005). Genericity from a cross-linguistic perspective. *Linguistics, 43*(2), 275-344.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Frieder, O., & Grossman, D. (2007). Temporal analysis of a very large topically categorized web query log. *Journal of the American Society for Information Science and Technology, 58*(2), 166-178.

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science, 5*, 133-143.

Belkin, N. J., Cool, C., Stein, A., & Thiel, U. (1995). Cases, scripts, and information-seeking strategies: On the design of interactive information retrieval systems. *Expert Systems with Applications, 9*, 379-395.

Belkin, N., & Croft, W. (1987). Retrieval techniques. *Annual Review of Information Science and Technology, 22*, 109-145.

Belkin, N., Cool, C., Kelly, D., Lee, H.-J., Muresan, G., Tang, M.-C., & Yuan, X.-J. (2003). Query length in interactive information retrieval. *SIGIR* (pp. 205-212). Toronto, Canada: ACM.

Bell, D. A. (2000). A formalism for relevance and its application in feature subset selection. *Machine Learning, 41*, 175-195.

Bergenholtz, H., & Gouws, R. (2010). A new perspective on the access process. *Hermes - Journal of Language and Communication Studies, 44*, pp. 103-128.

Bergsma, S., & Wang, Q. I. (2007). Learning noun phrase segmentation. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 819-826). Prague: Association for Computational Linguistics.

Bhavnani, S. K., Arbor, A., & Drabenstott, K. (2001). Towards a unified framework of IR tasks and strategies. *Proceedings of ASIST* (pp. 340-354). The American Society for Information Science and Technology.

Bhogal, J., Macfarlane, A., & Smith, P. (2007). A review of ontology based query expansion. *Information Processing & Management, 43*, 866-886.

Blair, D. C. (1980). Searching biases in large interactive document retrieval systems. *Journal of the American Society for Information Science, 31*(4), 271-277.

Blair, D. C. (2002a). The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size. *Information Processing & Management, 38*(2), 273-291.

Blair, D. C. (2002b). The challenge of commercial document retrieval, Part II: a strategy for document searching based on identifiable document partitions. *Information Processing & Management, 38*(2), 293-304.

Blair, D. C. (2003). Information retrieval and the philosophy of language. *Annual Review of Information Science and Technology, 37*(1), 3-50.

Blair, D. C. (2006). The data-document distinction revisited. *ACM SIGMIS Database, 37*(1), 77-96.

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM, 28*(3), 289-297.

Bloom, P., & Kelemen, D. (1995). Syntactic cues in the acquisition of collective nouns. *Cognition, 56*, 1-30.

Boeckx, C. (2010). *Language in Cognition.* Malden, MA/Oxford, UK: Wiley-Blackwell.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology, 54*(10), pp. 913-925.

Borlund, P., & Ingwersen, P. (1997). The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation, 53*(3), 225-250.

Brandt, C., Joachims, T., Yue, Y., & Bank, J. (2011). Dynamic ranked retrieval. *WSDM* (pp. 247-256). Hong Kong: ACM.

Brier, S. (1996). Cybersemiotics: A new interdisciplinary development applied to the problems of knowledge organisation and document retrieval in information science. *Journal of Documentation, 52*(3), 296-344.

Brier, S. (2008). *Cybersemiotics - Why information is not enough.* University of Toronto Press.

Broder, A. (2002). A taxonomy of web search. *ACM SIGIR Forum, 36*(2), pp. 3-10.

Brookes, B. (1977). The developing cognitive viewpoint in information science. In *CC-77: Int. Workshop on the Cognitive Viewpoint* (pp. 195-203). Ghent University.

Brookes, B. (1980). The foundations of information science. Part I: Philosophical aspects. *Journal of Information Science, 2*, 125-133.

Bruce, H. W. (1994). A cognitive view of the situational dynamism of user-centered relevance estimation. *Journal of the American Society for Information Science, 45*(3), pp. 142-148.

Bruner, J. (1973). *Beyond the information given: Studies in the psychology of knowing.* New York: Norton.

Bruner, J. (1986). *Actual minds, possible worlds.* Cambridge: Harvard University Press.

Brusilovsky, P. (2003). Adaptive navigation support in educational hypermedia: The role of student knowledge level and the case for meta-adaptation. *British Journal of Educational Technology, 34*(4), 487-497.

Buckland, M. K. (1991a). *Information and Information Systems.* London/New York: Greenwood.

Buckland, M. K. (1991b). Information as thing. *Journal of the American Society for Information Science, 42*(5), 351-360.

Buckland, M. K. (1997). What is a "document"? *Journal of the American Society for Information Science, 48*(9), 804-809.

Buckley, C., Salton, G., & Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. *SIGIR* (pp. 292-300). Dublin, Ireland: ACM.

Burnkrant, R. (1976). A motivational model of information-processing intensity. *Journal of Consumer Research, 3*, 21-30.

Callen, C. R. (2003 Mich. St. L. Rev. 1243). Rationality and relevancy: Conditional relevancy and constrained resources.

Chang, S.-J. L., & Lee, Y.-Y. (2000). Conceptualizing context and its relationship to the information behavior in dissertation research process. *Journal of Library and Information Science, 26*(2), 4-18.

Chang, Y., Ounis, I., & Kim, M. (2006). Query reformulation using automatically generated query concepts from a document space. *Information Processing & Management, 42*(2), 453-468.

Choi, Y., & Rasmussen, E. M. (2002). Users' relevance criteria in image retrieval in American history. *Information Processing & Management, 38*, 695-726.

Cleverdon, C. W. (1960). *ASLIB Cranfield Research Project - Report on the first stage of an investigation into the comparative efficiency of indexing systems.* Cranfield: The College of Aeronautics.

Comrie, B. (1985). *Tense.* Cambridge University Press.

Conrad, B. (1982). Referring and non-referring phrases - A study in the use of the gerund and the infinitive. *Publications of the Department of English, University of Copenhagen, 11*.

Cool, C. (1997). *Situation assessment in information retrieval interaction (Doctoral dissertation).* New Brunswick, NJ: Rutgers University.

Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval, 7*(1), 19-37.

Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science, 24*(2), 87-100.

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management, 36*, 533-550.

Croft, W., & Cruse, A. (2004). *Cognitive Linguistics.* Cambridge University Press.

Cuadra, C., & Katter, R. (1967). Opening the black box of 'relevance'. *Journal of Documentation, 23*(4), 291-303.

Dam, L., & Dam-Jensen, H. (2007). Instructions, context and compositionality: How to combine functional, lexical and interpretive dimensions of meaning. *Acta Linguistica Hafniensia, 39*, 83-100.

Dani, A. K. (2008). *Automatic identification of user goals in web search based on classification of click-through results (master's thesis).* Kharagpur, India: Dep. of Computer Science and Engineering, Indian Institute of Technology.

Daoud, M. (2009). *Accès personnalisé à l'information: approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche (Doctoral dissertation).* Toulouse: L'Université Paul Sabatier de Toulouse.

Daoud, M., Tamine-Lechani, L., & Boughanem, M. (2008). Learning user interests for a session-based personalized search. *IIiX* (pp. 57-64). London: ACM.

Daoud, M., Tamine-Lechani, L., & Boughanem, M. (2009). Towards a graph-based user profile modeling for a session-based personalized search. *Journal of Knowledge and Information Systems, 21*, pp. 365-398.

Daoud, M., Tamine-Lechani, L., Duy, D., & Boughanem, M. (2009). Contextual query classification for personalizing informational search. *WSDM.* Barcelona: ACM.

de Lima, E. F., & Pedersen, J. O. (1999). Phrase recognition and expansion for short, precision-biased queries based on a query log. *SIGIR* (pp. 145-152). Berkley, CA: ACM.

de Mey, M. (1980). The relevance of the cognitive paradigm for information science. In O. Harbo (Ed.), *Theory and Application of Information Research* (pp. 48-61). London: Mansell.

de Saussure, F. (1983). *Course in General Linguistics.* Peru, IL: Open Court.

Delphi Group. (2004). *Information Intelligence: Content Classification and the Enterprise Taxonomy Practice.* Boston, MA: Delphi Group.

Dervin, B. (2003). Chaos, order, and sense-making: A proposed theory for information design. In B. Dervin, & L. Foreman-Wernet (Eds.), *Sense-Making Methodology Reader: Selected writings of Brenda Dervin* (pp. 325-340). Cresskill, NJ: Hampton Press.

Dervin, B., & Dewdney, P. (1986). Neutral questioning: A new approach to the reference interview. *Research Quarterly, 25*, pp. 506-513.

Dervin, B., & Nilan, M. (1986). Information needs and uses. *ARIST, 21*, 3-33.

Dewey, J. (1933). *How we think.* Lexington: Heath.

Dey, A. (2001). Understanding and using context. *Personal and Ubiquitous Computing, 5*(1), 4-5.

Dik, S. C. (1997). *The Theory of Functional Grammar, vol. 1.* Berlin/New York: Mouton de Gruyter.

Doyle, L. B. (1963). Is relevance an adequate criterion in retrieval system evaluation? *Automation and Scientific Communication: Short Papers Contributed to the Theme Sessions of the 26th Annual Meeting of the American Documentation Institute.* Washington, D.C.: American Documentation Institute.

Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., & Robbins, D. (2003). Stuff I’ve seen: A system for personal information retrieval and re-use. *SIGIR* (pp. 72-79). Toronto, Canada: ACM.

Eggins, S. (2004). *An Introduction to Systemic Functional Linguistics* (2nd ed.). London: Continuum.

Elsweiler, D., Baillie, M., & Ruthven, I. (2011). What makes re-finding information difficult? - A study of email re-finding. *ECIR* (pp. 568-579). Dublin, Ireland: Springer Verlag.

Fauconnier, G. (1994). *Mental spaces - Aspects of meaning construction in natural language.* Cambridge University Press.

Fauconnier, G. (1997). *Mappings in Thought and Language.* Cambridge University Press.

Fidel, R. (1984). Online searching styles: A case-study-based model of searching behavior. *Journal of the American Society for Information Science, 35*(4), 211-221.

Flores Morador, F. (2012). Looking for Sartre: Remarks to a Phenomenology of Searching and Finding (unpubl. man., available from http://dx.doi.org/10.2139/ssrn.2179409).

Ford, N. (1980). Relating 'information needs' to learner characteristics in higher education. *Journal of Documentation, 36*, 99-114.

Gadamar, H.-G. (1960). *Wahrheit und Methode.* Tübingen: J.C.B. Mohr.

Gershenson, C. (2002). Complex Philosophy. *Proceedings of the 1st Biennial Seminar on Philosophical, Methodological & Epistemological Implications of Complexity Theory.* Habana, Cuba.

Gershenson, C. (2004). Cognitive paradigms: Which one is the best? *Cognitive Systems Research, 5*(2), 135-156.

Gershenson, C., & Heylighen, F. (2005). How can we think complex? - Philosophy, theory, and application. In K. A. Richardson (Ed.), *Managing Organizational Complexity* (pp. 47-61). Information Age Publishing.

Glover, E. J. (2001). *Using extra-topical user preferences to improve web-based metasearch (Ph.D. thesis).* The University of Michigan.

Grenon, P., & Smith, B. (2004). SNAP and SPAN: Towards dynamic spatial ontology. *Spatial Cognition and Computation, 4*(1), 69-103.

Grice, H. P. (1975). Logic and Conversation. In P. Cole, & J. Morgan, *Syntax and Semantics (vol. 3).* New York: Academic Press.

Grice, H. P. (1989). Logic and Conversation. In H. P. Grice, *Studies in the Way of Words* (pp. 22-40). Harvard University Press.

Grimes, C., Tang, D., & Russell, D. M. (2007). Query logs alone are not enough. *WWW.* Banff, Canada.

Gärdenfors, P. (1978). On the logic of relevance. *Synthese, 37*, 351-367.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science, 42*(1), 7-15.

Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43*(9), pp. 602-615.

Hattori, S., Tezuka, T., & Tanaka, K. (2007). Context-aware query refinement for mobile web search. *Proceedings of the 2007 international symposium on applications and the internet workshops* (p. 15). Washington, DC: IEEE Computer Society.

Hawkins, B. W. (1984). *The semantics of English spatial prepositions (Ph.D. thesis).* Univ. of California, San Diego.

Hearst, M. A. (1999). User Interfaces and Visualization. In R. Baeza-Yates, & B. Ribeiro-Neto (Eds.), *Modern Information Retrieval* (pp. 257-324). New York: Addison-Wesley.

Heidegger, M. (1996). *Being and Time - A translation of Sein und Zeit (orig. publ. 1953).* Albany: SUNY Press.

Hengeveld, K., & Mackenzie, J. L. (2008). *Functional Discourse Grammar - A typologically-based theory of language structure.* Oxford University Press.

Hertzum, M., Andersen, H. H., Andersen, V., & Hansen, C. B. (2002). Trust in information sources: Seeking information from people, documents, and virtual agents. *Interacting with Computers, 14*, 575-599.

Heylighen, F., & Joslyn, C. (2001). Cybernetics and Second-Order Cybernetics. In R. Meyers (Ed.), *Encyclopedia of Physical Science & Technology (3rd ed.).* New York: Academic Press.

Hjørland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science, 61*(2), 217-237.

Hjørland, B., & Christensen, F. S. (2002). Work tasks and socio-cognitive relevance: A specific example. *Journal of the American Society for Information Science and Technology, 53*(11), 960-965.

Ingwersen, P. (1982). Search procedures in the library analysed from the cognitive point of view. *Journal of Documentation, 38*, 165-191.

Ingwersen, P. (1992). *Information Retrieval Interaction.* London: Taylor Graham.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, 52*, 3-50.

Ingwersen, P. (2007). Context in information interaction – revisited 2006. *ProLISSA 2006: Proceedings of the Fourth Biennial DISSAnet Conference* (pp. 13-23). Pretoria: Infuse.

Ingwersen, P., & Järvelin, K. (2005a). *The turn: Integration of information seeking and retrieval in context.* Dordrecht: Springer Verlag.

Ingwersen, P., & Järvelin, K. (2005b). Information retrieval in context - IRiX. *SIGIR Forum, 38*(2), 31-39.

Ivanovas, G. (2010). *From Autism to Humanism - Contributions of system-theory to the understanding of therapy and health (doctoral dissertation).* The University of Crete.

Jackendoff, R. (1983). *Semantics and Cognition.* London: The MIT Press.

Jacsó, P. (2006). Pertinence in the eye of the user. *Online Information Review, 30*(1), 70-76.

Janes, J. W. (1994). Other people's judgments: A comparison of users' and others' judgments of document relevance, topicality, and utility. *Journal of the American Society for Information Science, 45*(3), 160-171.

Jansen, B., & Pooch, U. (2001). A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology, 52*(3), 235-246.

Jansen, B., Booth, D., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management, 44*(3), pp. 1251-1266.

Jansen, B., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the web. *SIGIR Forum* (pp. 5-17). ACM.

Jones, R., & Klinkner, K. (2008). Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. *CIKM* (pp. 699-708). New York: ACM.

Järvelin, K. (2011). IR research: Systems, interaction, evaluation and theories. *ECIR* (pp. 1-3). Dublin, Ireland: Springer-Verlag.

Kadmon, N., & Landman, F. (1993). Any. *Linguistics and Philosophy, 16*(4), 353-422.

Kajaba, M., & Navrat, P. (2009). Personalized web search using context enhanced query. *Proceedings of the International Conference on Computer Systems and Technologies.* ACM.

Kajaba, M., Navrat, P., & Chuda, D. (2009). A simple personalization layer improving relevancy of web search. *Computing and Information Systems Journal, 13*(3), pp. 29–35.

Keizer, M. E. (2005). Predicates as referring expressions. In P. Anstey, & J. Lachlan Mackenzie (Eds.), *Crucial Readings in Functional Grammar.* Berlin: Mouton de Gruyter.

Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology, 53*(13), 1120-1129.

Kellar, M. (2006). An examination of user behaviour during web information tasks. *CHI* (pp. 1763-1766). Montreal: ACM.

Kelly, D. (2006). Measuring online information seeking context, Part 1: Background and method. *Journal of the American Society for Information Science, 57*(13), 1729-1739.

Kelly, D., Dollu, V. D., & Fu, X. (2005). The loquacious user: A document-independent source of terms for query expansion. *SIGIR* (pp. 457-464). Salvador, Brazil: ACM.

Kelly, G. (1963). *A theory of personality: The psychology of personal constructs.* New York: Norton.

Kemp, D. (1974). Relevance, pertinence and information system development. *Information Storage and Retrieval, 10*(2), 37-47.

Kofler, C., Larson, M., & Hanjalic, A. (2011). To seek, perchance to fail: Expressions of user needs in internet video search. *ECIR* (pp. 611-616). Dublin, Ireland: Springer-Verlag.

Koschmann, T., Kuutti, K., & Hickman, L. (1998). The concept of breakdown in Heidegger, Leont'ev, and Dewey and its implications for education. *Mind, Culture, and Activity, 5*(1), 25-41.

Kraaij, W., & Pohlmann, R. (1996). *Using linguistic knowledge in information retrieval (UPLIFT technical report).*

Kuhlthau, C. C. (1993). A principle of uncertainty for information seeking. *Journal of Documentation, 49*(4), 339-355.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things - What categories reveal about the mind.* Chicago University Press.

Lambrecht, K. (1994). *Information structure and sentence form - Topic, focus and the mental representations of discourse referents.* Cambridge University Press.

Langacker, R. W. (2000). *Grammar and Conceptualization.* Berlin: Mouton de Gruyter.

Langacker, R. W. (2006). On the continuous debate about discreteness. *Cognitive Linguistics, 17*(1), 107-151.

Lee, J.-W., Kim, H.-J., & Lee, S.-G. (2010). Applying taxonomic knowledge and semantic collaborative filtering to personalized search: A bayesian belief network based approach. *12th International Asia-Pacific Web Conference*, (pp. 75-81).

Lee, U., Liu, Z., & Cho, J. (2005). Automatic identification of user goals in Web search. *WWW* (pp. 391-400). Chiba, Japan: ACM.

Leech, G. (1993). *Principles of Pragmatics.* London: Longman.

Li, C., Sun, A., & Datta, A. (2011). A generalized method for word sense disambiguation based on Wikipedia. *ECIR* (pp. 653-664). Dublin, Ireland: Springer-Verlag.

Li, X. (2010). Understanding the semantic structure of noun phrase queries. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1337-1345). Uppsala, Sweden: ACL.

Lopatovska, I., & Mokros, H. B. (2008). Willingness to pay and experienced utility as measures of affective value of information objects: Users' accounts. *Information Processing & Management, 44*, 92-104.

Luxenburger, J. (2008). *Modeling and Exploiting User Search Behavior for Information Retrieval (Doctoral dissertation).* Saarbrücken: Max-Planck-Institut für Informatik.

Lyons, J. (1977). *Semantics, vol. 2.* Cambridge University Press.

Machlup, F. (1983). Semantic quirks in studies of information. In F. Machlup, & U. Mansfield, *The Study of Information: Interdisciplinary Messages* (pp. 641-671). New York: Wiley.

Mackay, D. M. (1960). What makes the question? *The Listener, 63*, 789-790.

Maglaughlin, K. L., & Sonnenwald, D. H. (2002). User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the American Society for Information Science and Technology, 53*(5), 327-342.

Manshadi, M., & Li, X. (2009). Semantic tagging of Web search queries. *Proceedings of the 47th Annual Meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP* (pp. 861-869). Singapore: ACL/AFNLP.

Marchionini, G. (1995). *Information seeking in electronic environment.* Cambridge University Press.

Martins, D. S. (2009). *Uma abordagem para recuperacão de informacões sensível ao contexto usando retroalimantacão implícita de relevância (Ph.D. thesis).* São Carlos, Brazil.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., & Oltramari, A. (2003). *WonderWeb: Ontology Infrastructure for the Semantic Web - Deliverable D18: Ontology Library.* Trento, Italy: Laboratory For Applied Ontology - ISTC-CNR.

Maturana, H. R. (1978). Biology of Language: The epistemology of reality. In G. Miller, & E. Lenneberg (Eds.), *Psychology and Biology of Language and Thought: Essays in Honor of Eric Lenneberg* (pp. 27-64). New York: Academic Press.

Maturana, H. R. (1980). Biology of Cognition (reprint of 1970 paper). *Autopoiesis and Cognition: The Realization of the Living*, 5-58.

McKenzie, B., & Cockburn, A. (2001). An empirical analysis of web page revisitation. *HICSS.* Maui, Hawaii.

Merriam-Webster's Learner's Dictionary. (2012, November). Retrieved from http://www.learnersdictionary.com

Mithun, M. (1999). *The Languages of Native North America.* Cambridge University Press.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science, 48*(9), pp. 810-832.

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers, 10*, pp. 303-320.

Mooers, C. N. (1951). Zatocoding applied to mechanical organization of knowledge. *American Documentation, 2*, 20-32.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 56-60.

Morgan, C., & King, R. (1971). *Introduction to Psychology.* New York: McGraw-Hill.

Morris, R. C. (1994). Toward a user-centred information service. *Journal of the American Society for Information Science, 45*(1), pp. 20-30.

Murphy, G. L. (2002). *The Big Book of Concepts.* Cambridge, MA/London, UK: The MIT Press.

Navarro-Prieto, R., Scaife, M., & Rogers, Y. (1999). Cognitive strategies in Web searching. *5th Conference on human factors and the web.* Gaithersburg, Maryland.

O'Brien, P., Luo, X., Abou-Assaleh, T., Weizheng, G., & Li, S. (2009). Personalization of content ranking in the context of local search. *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies* (pp. 532-539). IEEE.

O'Connor. (1968). Some questions concerning "Information need". *American Documentation*, 200-204.

Olivera, P. A. (2010). *Specialised Dictionaries for Learners.* Walter de Gruyter.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Press.

Park, S., & Pennock, D. (2007). Applying collaborative filtering techniques to movie search for better ranking and browsing. *KDD*, (pp. 550-559). San José, CA.

Park, T. K. (1993). The nature of relevance in information retrieval: An empirical study. *The Library Quarterly, 63*(3), 318-351.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Francisco: Morgan Kaufmann Publishers, inc.

Peters, C., Oomen, J., Ibbotson, C., Ireson, N., Kamps, J., Jones, G., & Clough, P. (2006). *MultiMatch - Technology-enhanced Learning and Access to Cultural Heritage: State of the Art Report.* ISTI-CNR.

Pettigrew, K. E., & McKechnie, L. (2001). The use of theory in information science research. *Journal of the American Society for Information Science and Technology, 52*(1), 62-73.

Pharo, N., & Järvelin, K. (2004). The SST method: A tool for analysing Web information search processes. *Information Processing & Management, 40*(4), 633-654.

Phinitkar, P., & Sophatsathit, P. (2010). Personalization of search profile using ant foraging approach. *International Conference on Computation Science and its Applications (ICCSA)*, (pp. 209-224). Fukuoka, Japan.

Popovic, M., & Willett, P. (1992). The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science, 43*(5), 384-390.

Popper, K. (1972). *Objective Knowledge - An evolutionary approach.* Oxford: Clarendon Press.

Popper, K. (1978). *The Tanner Lecture on Human Values.* The University of Michigan.

Prinz, J. J. (2002). *Furnishing the Mind.* Cambridge, MA/London, UK: The MIT Press.

Rees, A. (1966). The relevance of relevance to the testing and evaluation of document retrieval systems. *ASLIB proceedings, 18*(11), 316-324.

Rees, A., & Schultz, D. (1967). *A field experiment approach to the study of relevance assessments in relation to document searching.* Cleveland, OH: Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University.

Riad, A. M., Eliminir, H. K., ElSoud, M. A., & Sabbeh, S. F. (2010). PSSE: An architecture for a personalized semantic search engine. *International Journal on Advances in Information Sciences and Service Sciences, 2*(1), pp. 102-112.

Riley, J. A. (2007). Mental representations: Reference and definiteness. *Journal of Pragmatics*, 831-871.

Robins, D. (2000). Interactive Information Retrieval: Context and basic notions. *Informing Science, 3*(2), 57-61.

Rose, D. E., & Levinson, D. (2004). Understanding user goals in web search. *WWW* (pp. 13-19). New York: ACM.

Rosenbaum, H. (1993). Information use environments and structuration: Towards an integration of Taylor and Giddens. *Proceedings of the 56th Annual Meeting of the American Society for Information Science*, (pp. 235-245).

Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. *SIGIR* (pp. 213-220). Toronto, Canada: ACM.

Salton, G., Allan, J., & Buckley, C. (1993). Approaches to passage retrieval in full text information systems. *SIGIR* (pp. 49-58). ACM.

Saracevic, T. (1970). The concept of "relevance" in information science: A historical review. In T. Saracevic, *Introduction to Information Science* (pp. 111-151). New York: R.R. Bowker.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science, 26*(6), 321-343.

Saracevic, T. (1996). Relevance reconsidered. *COLIS*, (pp. 201-208). Copenhagen, Denmark.

Saracevic, T. (1997a). Extension and application of the stratified model of information retrieval interaction. *Paper presented at the annual meeting of the American Society for Information Science*, *34*, pp. 313-327. Washington, D.C.

Saracevic, T. (1997b). The stratified model of information retrieval interaction: Extension and applications. *Proceedings of the American Society for Information Science, 34*, 313-327.

Saracevic, T. (1997c). Users lost: Reflections on the past, future, and limits of information science. *SIGIR Forum, 31*(2), 16-27.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology, 58*(13), 1915-1933.

Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving. II. Users, questions, and effectiveness. *Journal of the American Society for Information Science, 39*(3), 177-196.

Saracevic, T., Kantor, P., Chamis, A. Y., & Trivison, D. (1988). A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science, 39*(3), 161-176.

Sartre, J.-P. (1958). *Being and Nothingness.* Abingdon: Routledge.

Schamber, L. (1991). Users' criteria for evaluation in multimedia information seeking and use situations. *Proceedings of the 54th Annual Meeting of the American Society for lnformation Science*, *28*, pp. 126-133. Medford, NJ.

Schamber, L. (1996). What is a document? Rethinking the concept in uneasy times. *Journal of the American Society for Information Science, 47*(9), 669-671.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing & Management, 26*(6), 755-776.

Schneiderman, B., Byrd, D., & Croft, B. (1998). Sorting out searching: A user-interface framework for text searches. *Communications of the ACM, 41*(4), 95-98.

Schütz, A. (1970). *Reflections on the problem of relevance.* New Haven, CT: Yale University.

Schütz, A., & Luckmann, T. (1973). *The Structure of the Life-World.* Evanston, IL: Northwestern University Press.

Searle, J. R. (2001). Free will as a problem in neurobiology. *Philosophy, 76*, 491-514.

Sellen, A. J., Murphy, R., & Shaw, K. L. (2002). How knowledge workers use the web. *CHI* (pp. 227-234). Minneapolis, Minnesota: ACM.

Sieg, A., Mobasher, B., & Burke, R. (2007). Ontological user profiles for personalized web search. *Proceedings of the 5th Workshop on Intelligent Techniques for Web Personalization* (pp. 84-91). Vancouver, Canada: American Association for Artificial Intelligence.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum*, 6-12.

Simon, H. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, Communication, and the Public Interest.* Baltimore, MD: The Johns Hopkins Press.

Simon, H. A. (1976). *Administrative Behavior (3rd ed.).* New York: The Free Press.

Slobin, D. I. (2005). Linguistic representations of motion events - What is signifier and what is signified? In C. Maeder, O. Fischer, & W. J. Herlofsky (Eds.), *Outside - In, Inside - Out: Iconicity in Language and Litterature 4* (pp. 307-322). Amsterdam/Philadelphia: John Benjamins.

Smith, C. L., & Kantor, P. B. (2008). User adaptation: Good results from poor systems. *SIGIR* (pp. 147-154). Singapore: ACM.

Sonnenwald, D. (1999). Evolving perspectives of human information behavior: Contexts, situations, social networks and information horizons. *Exploring the Contexts of Information Behaviour: Proceedings of the 2nd International Conference in Information Needs, Use, and Seeking in Different Contexts (vol. 2).* London: Taylor Graham.

Sparck Jones, K. (1989, November). Retrieving information or answering questions? *British Library Annual Research Lecture*. London: The British Library.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition (M.A. thesis).* Cambridge: Harvard University Press.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind and Language, 17*(1/2), 3-23.

Spink, A., & Saracevic, T. (1998). Human-computer interaction in information retrieval: Nature and manifestations of feedback. *Interacting with Computers, 10*, 249-267.

Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. *Information Processing & Management, 34*, 599-621.

Spink, A., Ozmutlu, S., & Ozmutlu, H. C. (2002). U.S. versus European web searching trends. *SIGIR Forum, 36*(2), 32-38.

Stein, B., & Hagen, M. (2011). Introducing the user-over-ranking hypothesis. *ECIR* (pp. 503-509). Dublin, Ireland: Springer Verlag.

Swanson, D. R. (1977). Information retrieval as a trial-and-error process. *Library Quarterly, 47*, pp. 128-148.

Swanson, D. R. (1986). Subjective versus objective relevance in bibliographic retrieval systems. *Library Quarterly, 56*(4), 389-398.

Swanson, D. R. (1988). Historical Note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science, 39*(2), 92-98.

Sørensen, H. S. (1958). *Word-classes in Modern English with special reference to proper names - with an introductory theory of grammar, meaning and reference.* G.E.C. Gad.

Talmy, L. (1988). The relation of grammar to cognition. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics.* Amsterdam/Philadelphia: John Benjamins.

Tamine-Lechani, L., Boughanem, M., & Daoud, M. (2010). Evaluation of contextual information retrieval effectiveness: Overview of issues and research. *Knowledge and Information Systems, 24*, 1-34.

Tarp, S. (2008). The third leg of two-legged lexicography. *Hermes - Journal of Language and Communication Studies, 40*.

Tarp, S. (2009). Beyond lexicography: New visions and challenges in the information age. In H. Bergenholtz, S. Nielsen, & S. Tarp, *Lexicography at a crossroads* (pp. 17-32). Bern: Peter Lang.

Taylor, J. R. (2003). *Linguistic Categorization.* Oxford University Press.

Taylor, R. S. (1962). The process of asking questions. *American Documentation, 13*(4), 391-396.

Taylor, R. S. (1968). *Question-Negotiation and Information-Seeking in Libraries.* Bethlehem, Pennsylvania: Center for the Information Sciences, Lehigh University.

Taylor, R. S. (1993). Question-Negotiation and Information Seeking in Libraries. *College & Research Libraries, 29*, 178-194.

Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. *CHI* (pp. 415-422). Vienna, Austria: ACM.

Teevan, J., Dumais, S. T., & Liebling, D. J. (2008). To personalize or not to personalize: Modeling queries with variation in user intent. *SIGIR* (pp. 163-170). Singapore: ACM.

Terney, T. V. (2009). *The Combined Usage of Ontologies and Corpus Statistics in Information Retrieval (Ph.D. thesis).* Roskilde: Roskilde University.

Thayer, J. B. (1898). *A preliminary treatise on evidence at the common law.* Boston: Little, Brown, and Company.

Thomsen, O. N. (1992). Pronouns, word order, and prosody. Information structuring in preverbal slots in spoken Danish. *Copenhagen Working Papers in Linguistics, 2*, 119-232.

Tournadre, N., & Dorje, S. (2005). *Manual of Standard Tibetan.* Snow Lion.

Travis, C. (1981). *The True and the False - The Domain of the Pragmatic.* Amsterdam: John Benjamins Publishing Company.

Turpin, A. H., & Hersh, W. (2001). Why batch and user evaluations do not give the same results. *SIGIR* (pp. 225-231). New York: ACM.

Tversky, A., & Kahneman, D. (1974). Judgment and uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Vakkari, P., & Hakala, N. (2000). Changes in relevance criteria and problem stages in task performance. *Journal of Documentation, 56*, 540-562.

van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London: Butterworths.

van Rijsbergen, C. J. (1990). The science of information retrieval: Its methodology and logic. In *Conferentie Informatienwetenschap Nederland* (pp. 20-38). Haag: Rabin.

Verberne, S., Boves, L., & Kraaij, W. (2011). Bringing why-QA to web search. *ECIR* (pp. 491-496). Dublin, Ireland: Springer-Verlag.

Wang, J., & Zhu, J. (2009). Portfolio theory of information retrieval. *SIGIR* (pp. 115-122). Boston, MA: ACM.

Wang, P., Hawk, W. B., & Tenopir, C. (2000). Users' interaction with World Wide Web resources: An exploratory study using a holistic approach. *Information Processing & Management, 36*(2), 229-251.

Wilson, D., & Carston, R. (2007). A Unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In N. Burton-Roberts (Ed.), *Advances in Pragmatics* (pp. 230-260). Basingstoke: Palgrave Macmillan.

Wilson, D., & Sperber, D. (2004). Relevance Theory. In L. Horn, & G. Ward, *Handbook of Pragmatics.* Oxford: Blackwell.

Wilson, P. (1973). Situational Relevance. *Information Storage and Retrieval, 9*, pp. 457-471.

Wilson, T. (1981). On user studies and information needs. *Journal of Documentation, 37*(1), 3-15.

Wilson, T. (1997). Information behaviour: An interdisciplinary perspective. *Information Processing & Management, 33*(4), 551-572.

Winograd, T., & Flores, F. (1986). *Understanding Computers and Cognition - A new foundation for design.* Norwood, NJ: Ablex Publishing Corporation.

Wisniewski, E. J., Lamb, C. A., & Middleton, E. L. (2003). On the conceptual basis for the count and mass noun distinction. *Language and Cognitive Processes, 18*(5/6), 583-624.

Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., & Li, H. (2010). Context-aware ranking in web search. *SIGIR* (pp. 451-458). Geneva: ACM.

Xie, H. I. (2007). Shifts in information-seeking strategies in information retrieval in the digital age - A planned-situational model. *Information Research, 12*(4).

Xu, Y. C., & Chen, Z. (2006). Relevance Judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology, 57*(7), 961-973.

Yan, X., Lau, R. Y., Song, D., Li, X., & Ma, J. (2011). Toward a semantic granularity model for domain-specific information retrieval. *ACM Transactions on Information Systems, 29*(3), 1-46.

Zien, J., Meyer, J., & Tomlin, J. (2001). Web query characteristics and their implications on search engines. *WWW.* Hong Kong.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Cambridge, MA: Addison-Wesley.

TITLER I PH.D.SERIEN:

2004

1. Martin Grieger
Internet-based Electronic Marketplaces and Supply Chain Management

2. Thomas Basbøll
LIKENESS
A Philosophical Investigation

3. Morten Knudsen
Beslutningens vaklen
En systemteoretisk analyse of moderniseringen af et amtskommunalt sundhedsvæsen 1980-2000

4. Lars Bo Jeppesen
Organizing Consumer Innovation
A product development strategy that is based on online communities and allows some firms to benefit from a distributed process of innovation by consumers

5. Barbara Dragsted
SEGMENTATION IN TRANSLATION AND TRANSLATION MEMORY SYSTEMS
An empirical investigation of cognitive segmentation and effects of integrating a TM system into the translation process

6. Jeanet Hardis
Sociale partnerskaber
Et socialkonstruktivistisk casestudie af partnerskabsaktørers virkelighedsopfattelse mellem identitet og legitimitet

7. Henriette Hallberg Thygesen
System Dynamics in Action

8. Carsten Mejer Plath
Strategisk Økonomistyring

9. Annemette Kjærgaard
Knowledge Management as Internal Corporate Venturing
– a Field Study of the Rise and Fall of a Bottom-Up Process

10. Knut Arne Hovdal
De profesjonelle i endring
Norsk ph.d., ej til salg gennem Samfundslitteratur

11. Søren Jeppesen
Environmental Practices and Greening Strategies in Small Manufacturing Enterprises in South Africa
– A Critical Realist Approach

12. Lars Frode Frederiksen
Industriel forskningsledelse
– på sporet af mønstre og samarbejde i danske forskningsintensive virksomheder

13. Martin Jes Iversen
The Governance of GN Great Nordic
– in an age of strategic and structural transitions 1939-1988

14. Lars Pynt Andersen
The Rhetorical Strategies of Danish TV Advertising
A study of the first fifteen years with special emphasis on genre and irony

15. Jakob Rasmussen
Business Perspectives on E-learning

16. Sof Thrane
The Social and Economic Dynamics of Networks
– a Weberian Analysis of Three Formalised Horizontal Networks

17. Lene Nielsen
Engaging Personas and Narrative Scenarios – a study on how a user-centered approach influenced the perception of the design process in the e-business group at AstraZeneca

18. S.J Valstad
Organisationsidentitet
Norsk ph.d., ej til salg gennem Samfundslitteratur

19. Thomas Lyse Hansen
Six Essays on Pricing and Weather risk in Energy Markets

20. Sabine Madsen
Emerging Methods – An Interpretive Study of ISD Methods in Practice

21. Evis Sinani
The Impact of Foreign Direct Investment on Efficiency, Productivity Growth and Trade: An Empirical Investigation

22. Bent Meier Sørensen
Making Events Work Or,
How to Multiply Your Crisis

23. Pernille Schnoor
Brand Ethos
Om troværdige brand- og virksomhedsidentiteter i et retorisk og diskursteoretisk perspektiv

24. Sidsel Fabech
Von welchem Österreich ist hier die Rede?
Diskursive forhandlinger og magtkampe mellem rivaliserende nationale identitetskonstruktioner i østrigske pressediskurser

25. Klavs Odgaard Christensen
Sprogpolitik og identitetsdannelse i flersprogede forbundsstater
Et komparativt studie af Schweiz og Canada

26. Dana B. Minbaeva
Human Resource Practices and Knowledge Transfer in Multinational Corporations

27. Holger Højlund
Markedets politiske fornuft
Et studie af velfærdens organisering i perioden 1990-2003

28. Christine Mølgaard Frandsen
A.s erfaring
Om mellemværendets praktik i en transformation af mennesket og subjektiviteten

29. Sine Nørholm Just
The Constitution of Meaning
– A Meaningful Constitution?
Legitimacy, identity, and public opinion in the debate on the future of Europe

2005

1. Claus J. Varnes
Managing product innovation through rules – The role of formal and structured methods in product development

2. Helle Hedegaard Hein
Mellem konflikt og konsensus
– Dialogudvikling på hospitalsklinikker

3. Axel Rosenø
Customer Value Driven Product Innovation – A Study of Market Learning in New Product Development

4. Søren Buhl Pedersen
Making space
An outline of place branding

5. Camilla Funck Ellehave
Differences that Matter
An analysis of practices of gender and organizing in contemporary workplaces

6. Rigmor Madeleine Lond
Styring af kommunale forvaltninger

7. Mette Aagaard Andreassen
Supply Chain versus Supply Chain
Benchmarking as a Means to Managing Supply Chains

8. Caroline Aggestam-Pontoppidan
From an idea to a standard
The UN and the global governance of accountants' competence

9. Norsk ph.d.

10. Vivienne Heng Ker-ni
An Experimental Field Study on the

Effectiveness of Grocer Media Advertising
Measuring Ad Recall and Recognition, Purchase Intentions and Short-Term Sales

11. Allan Mortensen
Essays on the Pricing of Corporate Bonds and Credit Derivatives

12. Remo Stefano Chiari
Figure che fanno conoscere
Itinerario sull'idea del valore cognitivo e espressivo della metafora e di altri tropi da Aristotele e da Vico fino al cognitivismo contemporaneo

13. Anders McIlquham-Schmidt
Strategic Planning and Corporate Performance
An integrative research review and a meta-analysis of the strategic planning and corporate performance literature from 1956 to 2003

14. Jens Geersbro
The TDF – PMI Case
Making Sense of the Dynamics of Business Relationships and Networks

15 Mette Andersen
Corporate Social Responsibility in Global Supply Chains
Understanding the uniqueness of firm behaviour

16. Eva Boxenbaum
Institutional Genesis: Micro – Dynamic Foundations of Institutional Change

17. Peter Lund-Thomsen
Capacity Development, Environmental Justice NGOs, and Governance: The Case of South Africa

18. Signe Jarlov
Konstruktioner af offentlig ledelse

19. Lars Stæhr Jensen
Vocabulary Knowledge and Listening Comprehension in English as a Foreign Language
An empirical study employing data elicited from Danish EFL learners

20. Christian Nielsen
Essays on Business Reporting
Production and consumption of strategic information in the market for information

21. Marianne Thejls Fischer
Egos and Ethics of Management Consultants

22. Annie Bekke Kjær
Performance management i Proces-innovation
– belyst i et social-konstruktivistisk perspektiv

23. Suzanne Dee Pedersen
GENTAGELSENS METAMORFOSE
Om organisering af den kreative gøren i den kunstneriske arbejdspraksis

24. Benedikte Dorte Rosenbrink
Revenue Management
Økonomiske, konkurrencemæssige & organisatoriske konsekvenser

25. Thomas Riise Johansen
Written Accounts and Verbal Accounts
The Danish Case of Accounting and Accountability to Employees

26. Ann Fogelgren-Pedersen
The Mobile Internet: Pioneering Users' Adoption Decisions

27. Birgitte Rasmussen
Ledelse i fællesskab – de tillidsvalgtes fornyende rolle

28. Gitte Thit Nielsen
Remerger
– skabende ledelseskræfter i fusion og opkøb

29. Carmine Gioia
A MICROECONOMETRIC ANALYSIS OF MERGERS AND ACQUISITIONS

30. Ole Hinz
Den effektive forandringsleder: pilot, pædagog eller politiker?
Et studie i arbejdslederes meningstil-skrivninger i forbindelse med vellykket gennemførelse af ledelsesinitierede forandringsprojekter

31. Kjell-Åge Gotvassli
Et praksisbasert perspektiv på dynami-ske
læringsnettverk i toppidretten
Norsk ph.d., ej til salg gennem Samfundslitteratur

32. Henriette Langstrup Nielsen
Linking Healthcare
An inquiry into the changing perfor-mances of web-based technology for asthma monitoring

33. Karin Tweddell Levinsen
Virtuel Uddannelsespraksis
Master i IKT og Læring – et casestudie i hvordan proaktiv proceshåndtering kan forbedre praksis i virtuelle lærings-miljøer

34. Anika Liversage
Finding a Path
Labour Market Life Stories of Immigrant Professionals

35. Kasper Elmquist Jørgensen
Studier i samspillet mellem stat og erhvervsliv i Danmark under 1. verdenskrig

36. Finn Janning
A DIFFERENT STORY
Seduction, Conquest and Discovery

37. Patricia Ann Plackett
Strategic Management of the Radical Innovation Process
Leveraging Social Capital for Market Uncertainty Management

2006

1. Christian Vintergaard
Early Phases of Corporate Venturing

2. Niels Rom-Poulsen
Essays in Computational Finance

3. Tina Brandt Husman
Organisational Capabilities, Competitive Advantage & Project-Based Organisations
The Case of Advertising and Creative Good Production

4. Mette Rosenkrands Johansen
Practice at the top
– how top managers mobilise and use non-financial performance measures

5. Eva Parum
Corporate governance som strategisk kommunikations- og ledelsesværktøj

6. Susan Aagaard Petersen
Culture's Influence on Performance Management: The Case of a Danish Company in China

7. Thomas Nicolai Pedersen
The Discursive Constitution of Organi-zational Governance – Between unity and differentiation
The Case of the governance of environmental risks by World Bank environmental staff

8. Cynthia Selin
Volatile Visions: Transactons in Anticipatory Knowledge

9. Jesper Banghøj
Financial Accounting Information and Compensation in Danish Companies

10. Mikkel Lucas Overby
Strategic Alliances in Emerging High-Tech Markets: What's the Difference and does it Matter?

11. Tine Aage
External Information Acquisition of Industrial Districts and the Impact of Different Knowledge Creation Dimen-sions

A case study of the Fashion and Design Branch of the Industrial District of Montebelluna, NE Italy

12. Mikkel Flyverbom
Making the Global Information Society Governable
On the Governmentality of Multi-Stakeholder Networks

13. Anette Grønning
Personen bag
Tilstedevær i e-mail som inter-aktionsform mellem kunde og med-arbejder i dansk forsikringskontekst

14. Jørn Helder
One Company – One Language?
The NN-case

15. Lars Bjerregaard Mikkelsen
Differing perceptions of customer value
Development and application of a tool for mapping perceptions of customer value at both ends of customer-supplier dyads in industrial markets

16. Lise Granerud
Exploring Learning
Technological learning within small manufacturers in South Africa

17. Esben Rahbek Pedersen
Between Hopes and Realities: Reflections on the Promises and Practices of Corporate Social Responsibility (CSR)

18. Ramona Samson
The Cultural Integration Model and European Transformation.
The Case of Romania

2007

1. Jakob Vestergaard
Discipline in The Global Economy
Panopticism and the Post-Washington Consensus

2. Heidi Lund Hansen
Spaces for learning and working
A qualitative study of change of work, management, vehicles of power and social practices in open offices

3. Sudhanshu Rai
Exploring the internal dynamics of software development teams during user analysis
A tension enabled Institutionalization Model; "Where process becomes the objective"

4. Norsk ph.d.
Ej til salg gennem Samfundslitteratur

5. Serden Ozcan
EXPLORING HETEROGENEITY IN ORGANIZATIONAL ACTIONS AND OUTCOMES
A Behavioural Perspective

6. Kim Sundtoft Hald
Inter-organizational Performance Measurement and Management in Action
– An Ethnography on the Construction of Management, Identity and Relationships

7. Tobias Lindeberg
Evaluative Technologies
Quality and the Multiplicity of Performance

8. Merete Wedell-Wedellsborg
Den globale soldat
Identitetsdannelse og identitetsledelse i multinationale militære organisationer

9. Lars Frederiksen
Open Innovation Business Models
Innovation in firm-hosted online user communities and inter-firm project ventures in the music industry
– A collection of essays

10. Jonas Gabrielsen
Retorisk toposlære – fra statisk 'sted' til persuasiv aktivitet

11. Christian Moldt-Jørgensen
Fra meningsløs til meningsfuld evaluering.
Anvendelsen af studentertilfredshedsmålinger på de korte og mellemlange videregående uddannelser set fra et psykodynamisk systemperspektiv

12. Ping Gao
Extending the application of actor-network theory
Cases of innovation in the telecommunications industry

13. Peter Mejlby
Frihed og fængsel, en del af den samme drøm?
Et phronetisk baseret casestudie af frigørelsens og kontrollens sameksistens i værdibaseret ledelse!

14. Kristina Birch
Statistical Modelling in Marketing

15. Signe Poulsen
Sense and sensibility:
The language of emotional appeals in insurance marketing

16. Anders Bjerre Trolle
Essays on derivatives pricing and dynamic asset allocation

17. Peter Feldhütter
Empirical Studies of Bond and Credit Markets

18. Jens Henrik Eggert Christensen
Default and Recovery Risk Modeling and Estimation

19. Maria Theresa Larsen
Academic Enterprise: A New Mission for Universities or a Contradiction in Terms?
Four papers on the long-term implications of increasing industry involvement and commercialization in academia

20. Morten Wellendorf
Postimplementering af teknologi i den offentlige forvaltning
Analyser af en organisations kontinuerlige arbejde med informationsteknologi

21. Ekaterina Mhaanna
Concept Relations for Terminological Process Analysis

22. Stefan Ring Thorbjørnsen
Forsvaret i forandring
Et studie i officerers kapabiliteter under påvirkning af omverdenens forandringspres mod øget styring og læring

23. Christa Breum Amhøj
Det selvskabte medlemskab om managementstaten, dens styringsteknologier og indbyggere

24. Karoline Bromose
Between Technological Turbulence and Operational Stability
– An empirical case study of corporate venturing in TDC

25. Susanne Justesen
Navigating the Paradoxes of Diversity in Innovation Practice
– A Longitudinal study of six very different innovation processes – in practice

26. Luise Noring Henler
Conceptualising successful supply chain partnerships
– Viewing supply chain partnerships from an organisational culture perspective

27. Mark Mau
Kampen om telefonen
Det danske telefonvæsen under den tyske besættelse 1940-45

28. Jakob Halskov
The semiautomatic expansion of existing terminological ontologies using knowledge patterns discovered

on the WWW – an implementation and evaluation

29. Gergana Koleva
European Policy Instruments Beyond Networks and Structure: The Innovative Medicines Initiative

30. Christian Geisler Asmussen
Global Strategy and International Diversity: A Double-Edged Sword?

31. Christina Holm-Petersen
Stolthed og fordom
Kultur- og identitetsarbejde ved skabelsen af en ny sengeafdeling gennem fusion

32. Hans Peter Olsen
Hybrid Governance of Standardized States
Causes and Contours of the Global Regulation of Government Auditing

33. Lars Bøge Sørensen
Risk Management in the Supply Chain

34. Peter Aagaard
Det unikkes dynamikker
De institutionelle mulighedsbetingelser bag den individuelle udforskning i professionelt og frivilligt arbejde

35. Yun Mi Antorini
Brand Community Innovation
An Intrinsic Case Study of the Adult Fans of LEGO Community

36. Joachim Lynggaard Boll
Labor Related Corporate Social Performance in Denmark
Organizational and Institutional Perspectives

2008

1. Frederik Christian Vinten
Essays on Private Equity

2. Jesper Clement
Visual Influence of Packaging Design on In-Store Buying Decisions

3. Marius Brostrøm Kousgaard
Tid til kvalitetsmåling?
– Studier af indrulleringsprocesser i forbindelse med introduktionen af kliniske kvalitetsdatabaser i speciallægepraksissektoren

4. Irene Skovgaard Smith
Management Consulting in Action
Value creation and ambiguity in client-consultant relations

5. Anders Rom
Management accounting and integrated information systems
How to exploit the potential for management accounting of information technology

6. Marina Candi
Aesthetic Design as an Element of Service Innovation in New Technology-based Firms

7. Morten Schnack
Teknologi og tværfaglighed
– en analyse af diskussionen omkring indførelse af EPJ på en hospitalsafdeling

8. Helene Balslev Clausen
Juntos pero no revueltos – un estudio sobre emigrantes norteamericanos en un pueblo mexicano

9. Lise Justesen
Kunsten at skrive revisionsrapporter.
En beretning om forvaltningsrevisionens beretninger

10. Michael E. Hansen
The politics of corporate responsibility: CSR and the governance of child labor and core labor rights in the 1990s

11. Anne Roepstorff
Holdning for handling – en etnologisk undersøgelse af Virksomheders Sociale Ansvar/CSR

12. Claus Bajlum
Essays on Credit Risk and Credit Derivatives

13. Anders Bojesen
The Performative Power of Competence – an Inquiry into Subjectivity and Social Technologies at Work

14. Satu Reijonen
Green and Fragile
A Study on Markets and the Natural Environment

15. Ilduara Busta
Corporate Governance in Banking
A European Study

16. Kristian Anders Hvass
A Boolean Analysis Predicting Industry Change: Innovation, Imitation & Business Models
The Winning Hybrid: A case study of isomorphism in the airline industry

17. Trine Paludan
De uvidende og de udviklingsparate
Identitet som mulighed og restriktion blandt fabriksarbejdere på det aftayloriserede fabriksgulv

18. Kristian Jakobsen
Foreign market entry in transition economies: Entry timing and mode choice

19. Jakob Elming
Syntactic reordering in statistical machine translation

20. Lars Brømsøe Termansen
Regional Computable General Equilibrium Models for Denmark
Three papers laying the foundation for regional CGE models with agglomeration characteristics

21. Mia Reinholt
The Motivational Foundations of Knowledge Sharing

22. Frederikke Krogh-Meibom
The Co-Evolution of Institutions and Technology
– A Neo-Institutional Understanding of Change Processes within the Business Press – the Case Study of Financial Times

23. Peter D. Ørberg Jensen
OFFSHORING OF ADVANCED AND HIGH-VALUE TECHNICAL SERVICES: ANTECEDENTS, PROCESS DYNAMICS AND FIRMLEVEL IMPACTS

24. Pham Thi Song Hanh
Functional Upgrading, Relational Capability and Export Performance of Vietnamese Wood Furniture Producers

25. Mads Vangkilde
Why wait?
An Exploration of first-mover advantages among Danish e-grocers through a resource perspective

26. Hubert Buch-Hansen
Rethinking the History of European Level Merger Control
A Critical Political Economy Perspective

2009

1. Vivian Lindhardsen
From Independent Ratings to Communal Ratings: A Study of CWA Raters' Decision-Making Behaviours

2. Guðrið Weihe
Public-Private Partnerships: Meaning and Practice

3. Chris Nøkkentved
Enabling Supply Networks with Collaborative Information Infrastructures
An Empirical Investigation of Business Model Innovation in Supplier Relationship Management

4. Sara Louise Muhr
Wound, Interrupted – On the Vulnerability of Diversity Management

5. Christine Sestoft
Forbrugeradfærd i et Stats- og Livsformsteoretisk perspektiv

6. Michael Pedersen
Tune in, Breakdown, and Reboot: On the production of the stress-fit self-managing employee

7. Salla Lutz
Position and Reposition in Networks – Exemplified by the Transformation of the Danish Pine Furniture Manufacturers

8. Jens Forssbæck
Essays on market discipline in commercial and central banking

9. Tine Murphy
Sense from Silence – A Basis for Organised Action
How do Sensemaking Processes with Minimal Sharing Relate to the Reproduction of Organised Action?

10. Sara Malou Strandvad
Inspirations for a new sociology of art: A sociomaterial study of development processes in the Danish film industry

11. Nicolaas Mouton
On the evolution of social scientific metaphors:
A cognitive-historical enquiry into the divergent trajectories of the idea that collective entities – states and societies, cities and corporations – are biological organisms.

12. Lars Andreas Knutsen
Mobile Data Services:
Shaping of user engagements

13. Nikolaos Theodoros Korfiatis
Information Exchange and Behavior
A Multi-method Inquiry on Online Communities

14. Jens Albæk
Forestillinger om kvalitet og tværfaglighed på sygehuse
– skabelse af forestillinger i læge- og plejegrupperne angående relevans af nye idéer om kvalitetsudvikling gennem tolkningsprocesser

15. Maja Lotz
The Business of Co-Creation – and the Co-Creation of Business

16. Gitte P. Jakobsen
Narrative Construction of Leader Identity in a Leader Development Program Context

17. Dorte Hermansen
"Living the brand" som en brandorienteret dialogisk praxis:
Om udvikling af medarbejdernes brandorienterede dømmekraft

18. Aseem Kinra
Supply Chain (logistics) Environmental Complexity

19. Michael Nørager
How to manage SMEs through the transformation from non innovative to innovative?

20. Kristin Wallevik
Corporate Governance in Family Firms
The Norwegian Maritime Sector

21. Bo Hansen Hansen
Beyond the Process
Enriching Software Process Improvement with Knowledge Management

22. Annemette Skot-Hansen
Franske adjektivisk afledte adverbier, der tager præpositionssyntagmer indledt med præpositionen à som argumenter
En valensgrammatisk undersøgelse

23. Line Gry Knudsen
Collaborative R&D Capabilities
In Search of Micro-Foundations

24. Christian Scheuer
Employers meet employees
Essays on sorting and globalization

25. Rasmus Johnsen
The Great Health of Melancholy
A Study of the Pathologies of Performativity

26. Ha Thi Van Pham
Internationalization, Competitiveness Enhancement and Export Performance of Emerging Market Firms:
Evidence from Vietnam

27. Henriette Balieu
Kontrolbegrebets betydning for kausativalternationen i spansk
En kognitiv-typologisk analyse

2010

1. Yen Tran
Organizing Innovationin Turbulent Fashion Market
Four papers on how fashion firms create and appropriate innovation value

2. Anders Raastrup Kristensen
Metaphysical Labour
Flexibility, Performance and Commitment in Work-Life Management

3. Margrét Sigrún Sigurdardottir
Dependently independent
Co-existence of institutional logics in the recorded music industry

4. Ásta Dis Óladóttir
Internationalization from a small domestic base:
An empirical analysis of Economics and Management

5. Christine Secher
E-deltagelse i praksis – politikernes og forvaltningens medkonstruktion og konsekvenserne heraf

6. Marianne Stang Våland
What we talk about when we talk about space:
End User Participation between Processes of Organizational and Architectural Design

7. Rex Degnegaard
Strategic Change Management
Change Management Challenges in the Danish Police Reform

8. Ulrik Schultz Brix
Værdi i rekruttering – den sikre beslutning
En pragmatisk analyse af perception og synliggørelse af værdi i rekrutterings- og udvælgelsesarbejdet

9. Jan Ole Similä
Kontraktsledelse
Relasjonen mellom virksomhetsledelse og kontraktshåndtering, belyst via fire norske virksomheter

10. Susanne Boch Waldorff
Emerging Organizations: In between local translation, institutional logics and discourse

11. Brian Kane
Performance Talk
Next Generation Management of Organizational Performance

12. Lars Ohnemus
Brand Thrust: Strategic Branding and Shareholder Value
An Empirical Reconciliation of two Critical Concepts

13. Jesper Schlamovitz
Håndtering af usikkerhed i film- og byggeprojekter

14. Tommy Moesby-Jensen
Det faktiske livs forbindtlighed
Førsokratisk informeret, ny-aristotelisk ἦθος-tænkning hos Martin Heidegger

15. Christian Fich
Two Nations Divided by Common Values
French National Habitus and the Rejection of American Power

16. Peter Beyer
Processer, sammenhængskraft og fleksibilitet
Et empirisk casestudie af omstillingsforløb i fire virksomheder

17. Adam Buchhorn
Markets of Good Intentions
Constructing and Organizing Biogas Markets Amid Fragility and Controversy

18. Cecilie K. Moesby-Jensen
Social læring og fælles praksis
Et mixed method studie, der belyser læringskonsekvenser af et lederkursus for et praksisfællesskab af offentlige mellemledere

19. Heidi Boye
Fødevarer og sundhed i senmodernismen
– En indsigt i hyggefænomenet og de relaterede fødevarepraksisser

20. Kristine Munkgård Pedersen
Flygtige forbindelser og midlertidige mobiliseringer
Om kulturel produktion på Roskilde Festival

21. Oliver Jacob Weber
Causes of Intercompany Harmony in Business Markets – An Empirical Investigation from a Dyad Perspective

22. Susanne Ekman
Authority and Autonomy
Paradoxes of Modern Knowledge Work

23. Anette Frey Larsen
Kvalitetsledelse på danske hospitaler
– Ledelsernes indflydelse på introduktion og vedligeholdelse af kvalitetsstrategier i det danske sundhedsvæsen

24. Toyoko Sato
Performativity and Discourse: Japanese Advertisements on the Aesthetic Education of Desire

25. Kenneth Brinch Jensen
Identifying the Last Planner System
Lean management in the construction industry

26. Javier Busquets
Orchestrating Network Behavior for Innovation

27. Luke Patey
The Power of Resistance: India's National Oil Company and International Activism in Sudan

28. Mette Vedel
Value Creation in Triadic Business Relationships. Interaction, Interconnection and Position

29. Kristian Tørning
Knowledge Management Systems in Practice – A Work Place Study

30. Qingxin Shi
An Empirical Study of Thinking Aloud Usability Testing from a Cultural Perspective

31. Tanja Juul Christiansen
Corporate blogging: Medarbejderes kommunikative handlekraft

32. Malgorzata Ciesielska
Hybrid Organisations.
A study of the Open Source – business setting

33. Jens Dick-Nielsen
Three Essays on Corporate Bond Market Liquidity

34. Sabrina Speiermann
Modstandens Politik
Kampagnestyring i Velfærdsstaten.
En diskussion af trafikkampagners styringspotentiale

35. Julie Uldam
Fickle Commitment. Fostering political engagement in 'the flighty world of online activism'

36. Annegrete Juul Nielsen
Traveling technologies and transformations in health care

37. Athur Mühlen-Schulte
Organising Development
Power and Organisational Reform in the United Nations Development Programme

38. Louise Rygaard Jonas
Branding på butiksgulvet
Et case-studie af kultur- og identitets-arbejdet i Kvickly

2011

1. Stefan Fraenkel
Key Success Factors for Sales Force Readiness during New Product Launch
A Study of Product Launches in the Swedish Pharmaceutical Industry

2. Christian Plesner Rossing
International Transfer Pricing in Theory and Practice

3. Tobias Dam Hede
Samtalekunst og ledelsesdisciplin
– en analyse af coachingsdiskursens genealogi og governmentality

4. Kim Pettersson
Essays on Audit Quality, Auditor Choice, and Equity Valuation

5. Henrik Merkelsen
The expert-lay controversy in risk research and management. Effects of institutional distances. Studies of risk definitions, perceptions, management and communication

6. Simon S. Torp
Employee Stock Ownership:
Effect on Strategic Management and Performance

7. Mie Harder
Internal Antecedents of Management Innovation

8. Ole Helby Petersen
Public-Private Partnerships: Policy and Regulation – With Comparative and Multi-level Case Studies from Denmark and Ireland

9. Morten Krogh Petersen
'Good' Outcomes. Handling Multiplicity in Government Communication

10. Kristian Tangsgaard Hvelplund
Allocation of cognitive resources in translation - an eye-tracking and key-logging study

11. Moshe Yonatany
The Internationalization Process of Digital Service Providers

12. Anne Vestergaard
Distance and Suffering
Humanitarian Discourse in the age of Mediatization

13. Thorsten Mikkelsen
Personligsheds indflydelse på forretningsrelationer

14. Jane Thostrup Jagd
Hvorfor fortsætter fusionsbølgen udover "the tipping point"?
– en empirisk analyse af information og kognitioner om fusioner

15. Gregory Gimpel
Value-driven Adoption and Consumption of Technology: Understanding Technology Decision Making

16. Thomas Stengade Sønderskov
Den nye mulighed
Social innovation i en forretningsmæssig kontekst

17. Jeppe Christoffersen
Donor supported strategic alliances in developing countries

18. Vibeke Vad Baunsgaard
Dominant Ideological Modes of Rationality: Cross functional

integration in the process of product innovation

19. Throstur Olaf Sigurjonsson
Governance Failure and Icelands's Financial Collapse

20. Allan Sall Tang Andersen
Essays on the modeling of risks in interest-rate and inflation markets

21. Heidi Tscherning
Mobile Devices in Social Contexts

22. Birgitte Gorm Hansen
Adapting in the Knowledge Economy
Lateral Strategies for Scientists and Those Who Study Them

23. Kristina Vaarst Andersen
Optimal Levels of Embeddedness
The Contingent Value of Networked Collaboration

24. Justine Grønbæk Pors
Noisy Management
A History of Danish School Governing from 1970-2010

25. Stefan Linder
Micro-foundations of Strategic Entrepreneurship
Essays on Autonomous Strategic Action

26. Xin Li
Toward an Integrative Framework of National Competitiveness
An application to China

27. Rune Thorbjørn Clausen
Værdifuld arkitektur
Et eksplorativt studie af bygningers rolle i virksomheders værdiskabelse

28. Monica Viken
Markedsundersøkelser som bevis i varemerke- og markedsføringsrett

29. Christian Wymann
Tattooing
The Economic and Artistic Constitution of a Social Phenomenon

30. Sanne Frandsen
Productive Incoherence
A Case Study of Branding and Identity Struggles in a Low-Prestige Organization

31. Mads Stenbo Nielsen
Essays on Correlation Modelling

32. Ivan Häuser
Følelse og sprog
Etablering af en ekspressiv kategori, eksemplificeret på russisk

33. Sebastian Schwenen
Security of Supply in Electricity Markets

2012

1. Peter Holm Andreasen
The Dynamics of Procurement Management
- A Complexity Approach

2. Martin Haulrich
Data-Driven Bitext Dependency Parsing and Alignment

3. Line Kirkegaard
Konsulenten i den anden nat
En undersøgelse af det intense arbejdsliv

4. Tonny Stenheim
Decision usefulness of goodwill under IFRS

5. Morten Lind Larsen
Produktivitet, vækst og velfærd
Industrirådet og efterkrigstidens Danmark 1945 - 1958

6. Petter Berg
Cartel Damages and Cost Asymmetries

7. Lynn Kahle
Experiential Discourse in Marketing
A methodical inquiry into practice and theory

8. Anne Roelsgaard Obling
Management of Emotions in Accelerated Medical Relationships

9. Thomas Frandsen
Managing Modularity of Service Processes Architecture

10. Carina Christine Skovmøller
*CSR som noget særligt
Et casestudie om styring og menings-skabelse i relation til CSR ud fra en intern optik*

11. Michael Tell
*Fradragsbeskæring af selskabers finansieringsudgifter
En skatteretlig analyse af SEL §§ 11, 11B og 11C*

12. Morten Holm
*Customer Profitability Measurement Models
Their Merits and Sophistication across Contexts*

13. Katja Joo Dyppel
*Beskatning af derivater
En analyse af dansk skatteret*

14. Esben Anton Schultz
*Essays in Labor Economics
Evidence from Danish Micro Data*

15. Carina Risvig Hansen
"Contracts not covered, or not fully covered, by the Public Sector Directive"

16. Anja Svejgaard Pors
*Iværksættelse af kommunikation
- patientfigurer i hospitalets strategiske kommunikation*

17. Frans Bévort
*Making sense of management with logics
An ethnographic study of accountants who become managers*

18. René Kallestrup
The Dynamics of Bank and Sovereign Credit Risk

19. Brett Crawford
*Revisiting the Phenomenon of Interests in Organizational Institutionalism
The Case of U.S. Chambers of Commerce*

20. Mario Daniele Amore
Essays on Empirical Corporate Finance

21. Arne Stjernholm Madsen
*The evolution of innovation strategy
Studied in the context of medical device activities at the pharmaceutical company Novo Nordisk A/S in the period 1980-2008*

22. Jacob Holm Hansen
*Is Social Integration Necessary for Corporate Branding?
A study of corporate branding strategies at Novo Nordisk*

23. Stuart Webber
Corporate Profit Shifting and the Multinational Enterprise

24. Helene Ratner
*Promises of Reflexivity
Managing and Researching Inclusive Schools*

25. Therese Strand
The Owners and the Power: Insights from Annual General Meetings

26. Robert Gavin Strand
In Praise of Corporate Social Responsibility Bureaucracy

27. Nina Sormunen
*Auditor's going-concern reporting
Reporting decision and content of the report*

28. John Bang Mathiasen
*Learning within a product development working practice:
- an understanding anchored in pragmatism*

29. Philip Holst Riis
Understanding Role-Oriented Enterprise Systems: From Vendors to Customers

30. Marie Lisa Dacanay
*Social Enterprises and the Poor
Enhancing Social Entrepreneurship and Stakeholder Theory*

31. Fumiko Kano Glückstad
Bridging Remote Cultures: Cross-lingual concept mapping based on the information receiver's prior-knowledge

32. Henrik Barslund Fosse
Empirical Essays in International Trade

33. Peter Alexander Albrecht
Foundational hybridity and its reproduction
Security sector reform in Sierra Leone

34. Maja Rosenstock
CSR - hvor svært kan det være?
Kulturanalytisk casestudie om udfordringer og dilemmaer med at forankre Coops CSR-strategi

35. Jeanette Rasmussen
Tweens, medier og forbrug
Et studie af 10-12 årige danske børns brug af internettet, opfattelse og forståelse af markedsføring og forbrug

36. Ib Tunby Gulbrandsen
'This page is not intended for a US Audience'
A five-act spectacle on online communication, collaboration & organization.

37. Kasper Aalling Teilmann
Interactive Approaches to Rural Development

38. Mette Mogensen
The Organization(s) of Well-being and Productivity
(Re)assembling work in the Danish Post

39. Søren Friis Møller
From Disinterestedness to Engagement Towards Relational Leadership In the Cultural Sector

40. Nico Peter Berhausen
Management Control, Innovation and Strategic Objectives – Interactions and Convergence in Product Development Networks

41. Balder Onarheim
Creativity under Constraints
Creativity as Balancing 'Constrainedness'

42. Haoyong Zhou
Essays on Family Firms

43. Elisabeth Naima Mikkelsen
Making sense of organisational conflict
An empirical study of enacted sensemaking in everyday conflict at work

2013

1. Jacob Lyngsie
Entrepreneurship in an Organizational Context

2. Signe Groth-Brodersen
Fra ledelse til selvet
En socialpsykologisk analyse af forholdet imellem selvledelse, ledelse og stress i det moderne arbejdsliv

3. Nis Høyrup Christensen
Shaping Markets: A Neoinstitutional Analysis of the Emerging Organizational Field of Renewable Energy in China

4. Christian Edelvold Berg
As a matter of size
THE IMPORTANCE OF CRITICAL MASS AND THE CONSEQUENCES OF SCARCITY FOR TELEVISION MARKETS

5. Christine D. Isakson
Coworker Influence and Labor Mobility
Essays on Turnover, Entrepreneurship and Location Choice in the Danish Maritime Industry

6. Niels Joseph Jerne Lennon
Accounting Qualities in Practice
Rhizomatic stories of representational faithfulness, decision making and control

7. Shannon O'Donnell
Making Ensemble Possible
How special groups organize for collaborative creativity in conditions of spatial variability and distance

8. Robert W. D. Veitch
Access Decisions in a Partly-Digital World Comparing Digital Piracy and Legal Modes for Film and Music

9. Marie Mathiesen
Making Strategy Work An Organizational Ethnography

10. Arisa Shollo
The role of business intelligence in organizational decision-making

11. Mia Kaspersen
The construction of social and environmental reporting

12. Marcus Møller Larsen
The organizational design of offshoring

13. Mette Ohm Rørdam
EU Law on Food Naming The prohibition against misleading names in an internal market context

14. Hans Peter Rasmussen
GIV EN GED! Kan giver-idealtyper forklare støtte til velgørenhed og understøtte relationsopbygning?

15. Ruben Schachtenhaufen
Fonetisk reduktion i dansk

16. Peter Koerver Schmidt
Dansk CFC-beskatning I et internationalt og komparativt perspektiv

17. Morten Froholdt
Strategi i den offentlige sektor En kortlægning af styringsmæssig kontekst, strategisk tilgang, samt anvendte redskaber og teknologier for udvalgte danske statslige styrelser

18. Annette Camilla Sjørup
Cognitive effort in metaphor translation An eye-tracking and key-logging study

19. Tamara Stucchi
The Internationalization of Emerging Market Firms: A Context-Specific Study

20. Thomas Lopdrup-Hjorth
"Let's Go Outside": The Value of Co-Creation

21. Ana Alačovska
Genre and Autonomy in Cultural Production The case of travel guidebook production

22. Marius Gudmand-Høyer
Stemningssindssygdommenes historie i det 19. århundrede Omtydningen af melankolien og manien som bipolære stemningslidelser i dansk sammenhæng under hensyn til dannelsen af det moderne følelseslivs relative autonomi. En problematiserings- og erfarings-analytisk undersøgelse

23. Lichen Alex Yu
Fabricating an S&OP Process Circulating References and Matters of Concern

24. Esben Alfort
The Expression of a Need Understanding search

TITLER I ATV PH.D.-SERIEN

1992

1. Niels Kornum
Servicesamkørsel – organisation, økonomi og planlægningsmetode

1995

2. Verner Worm
Nordiske virksomheder i Kina
Kulturspecifikke interaktionsrelationer ved nordiske virksomhedsetableringer i Kina

1999

3. Mogens Bjerre
Key Account Management of Complex Strategic Relationships
An Empirical Study of the Fast Moving Consumer Goods Industry

2000

4. Lotte Darsø
Innovation in the Making
Interaction Research with heterogeneous Groups of Knowledge Workers creating new Knowledge and new Leads

2001

5. Peter Hobolt Jensen
Managing Strategic Design Identities
The case of the Lego Developer Network

2002

6. Peter Lohmann
The Deleuzian Other of Organizational Change – Moving Perspectives of the Human

7. Anne Marie Jess Hansen
To lead from a distance: The dynamic interplay between strategy and strategizing – A case study of the strategic management process

2003

8. Lotte Henriksen
Videndeling
– om organisatoriske og ledelsesmæssige udfordringer ved videndeling i praksis

9. Niels Christian Nickelsen
Arrangements of Knowing: Coordinating Procedures Tools and Bodies in Industrial Production – a case study of the collective making of new products

2005

10. Carsten Ørts Hansen
Konstruktion af ledelsesteknologier og effektivitet

TITLER I DBA PH.D.-SERIEN

2007

1. Peter Kastrup-Misir
Endeavoring to Understand Market Orientation – and the concomitant co-mutation of the researched, the re searcher, the research itself and the truth

2009

1. Torkild Leo Thellefsen
Fundamental Signs and Significance effects
A Semeiotic outline of Fundamental Signs, Significance-effects, Knowledge Profiling and their use in Knowledge Organization and Branding

2. Daniel Ronzani
When Bits Learn to Walk Don’t Make Them Trip. Technological Innovation and the Role of Regulation by Law in Information Systems Research: the Case of Radio Frequency Identification (RFID)

2010

1. Alexander Carnera
Magten over livet og livet som magt
Studier i den biopolitiske ambivalens