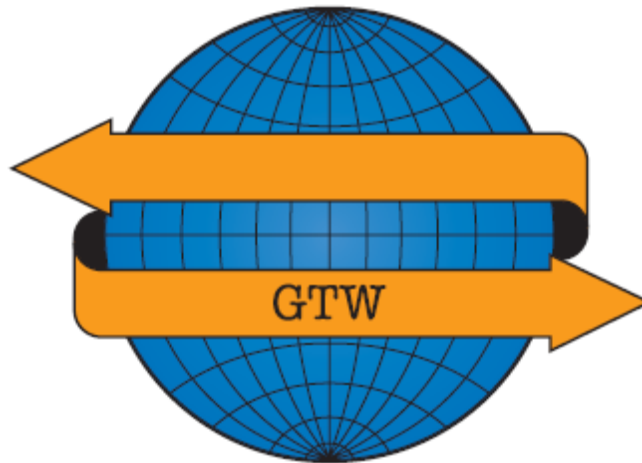


TERM BASES AND LINGUISTIC LINKED OPEN DATA

Edited by:
Hanne Erdman Thomsen,
Antonio Pareja-Lora &
Bodil Nistrup Madsen



TKE 2016

12th International conference on
Terminology and Knowledge Engineering



TERM BASES AND LINGUISTIC LINKED OPEN DATA

Copenhagen Business School

Department of International Business Communication

<https://sf.cbs.dk/gtw>

ISBN 978-87-999179-0-7

Preface

This volume of proceedings contains 17 papers to be presented at the 12th International Conference on Terminology and Knowledge Engineering, TKE 2016. 22 full papers were submitted in total, and they were all peer-reviewed by at least two reviewers with knowledge within the specific subfield as indicated by the authors. As it can be easily inferred, not all papers submitted could be presented at the conference (the acceptance ratio in this edition of TKE is 77%), but it is our hope that the authors of the papers eventually rejected will find the comments received from the reviewers useful for pursuing their work.

The theme of this year's TKE is 'Term Bases and Linguistic Linked Open Data'.

Mono- and multi-lingual term bases, which contain information about concepts (terms, definitions, examples of use, references, comments on equivalence etc.), have always made up valuable linguistic resources. Today, some terminology and knowledge bases combine traditional term bases and terminological ontologies, where concepts are related by means of various types of concept relations, and are further described by means of characteristics.

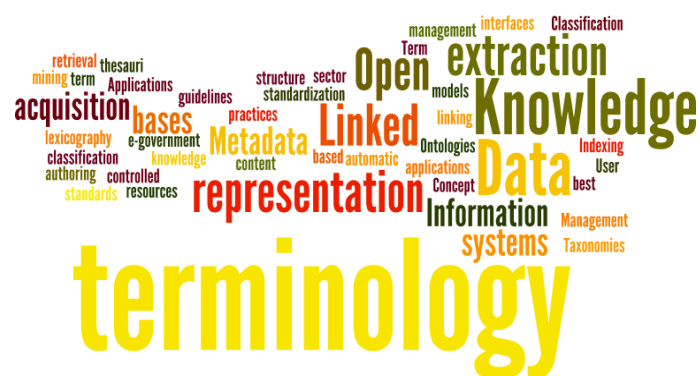
Besides, there is a new trend to represent knowledge as linked data, a new knowledge representation formalism in which data are structured as a network (or a set of networks) of resources, and in which the focus is rather on the relations between resources and their particular properties. Most frequently, linked data are published as linked open data, that is, data that are freely accessible and reusable. And, most interestingly, also linguistic (as well as terminological) knowledge has also been represented more recently as linked data.

Accordingly, some of the main aims of TKE 2016 will be

- to discuss the theories, best practices, guidelines, methods, techniques and tools developed for terminology and knowledge bases (including data and/or knowledge structure and acquisition, validation of knowledge, information and data, as well as user interfaces),
- to compare these with the theories, best practices, guidelines, methods, techniques and tools developed in the framework of the Linguistic Linked Open Data initiatives,
- and to identify actual and potential synergies, complementarities, and divergences between these two research and development areas.

On the one hand, the comprehensive list of topics addressed in the call for papers is shown in Figure 1. Clearly, the main potential topics of interest included the terms 'terminology', 'knowledge', 'extraction', 'representation', 'linked [open] data', 'metadata' and 'information'.

Fig. 1. Topics addressed in the call for papers for TKE 2016.



On the other hand, the topics addressed in the accepted papers are summarized in Figure 2. The main actual topics covered by these 17 papers can be summarized by the following terms: ‘terminology’, ‘management’, ‘concept [structure]’, ‘knowledge’, ‘corpus’, ‘system’, and ‘[semantic] annotation’.

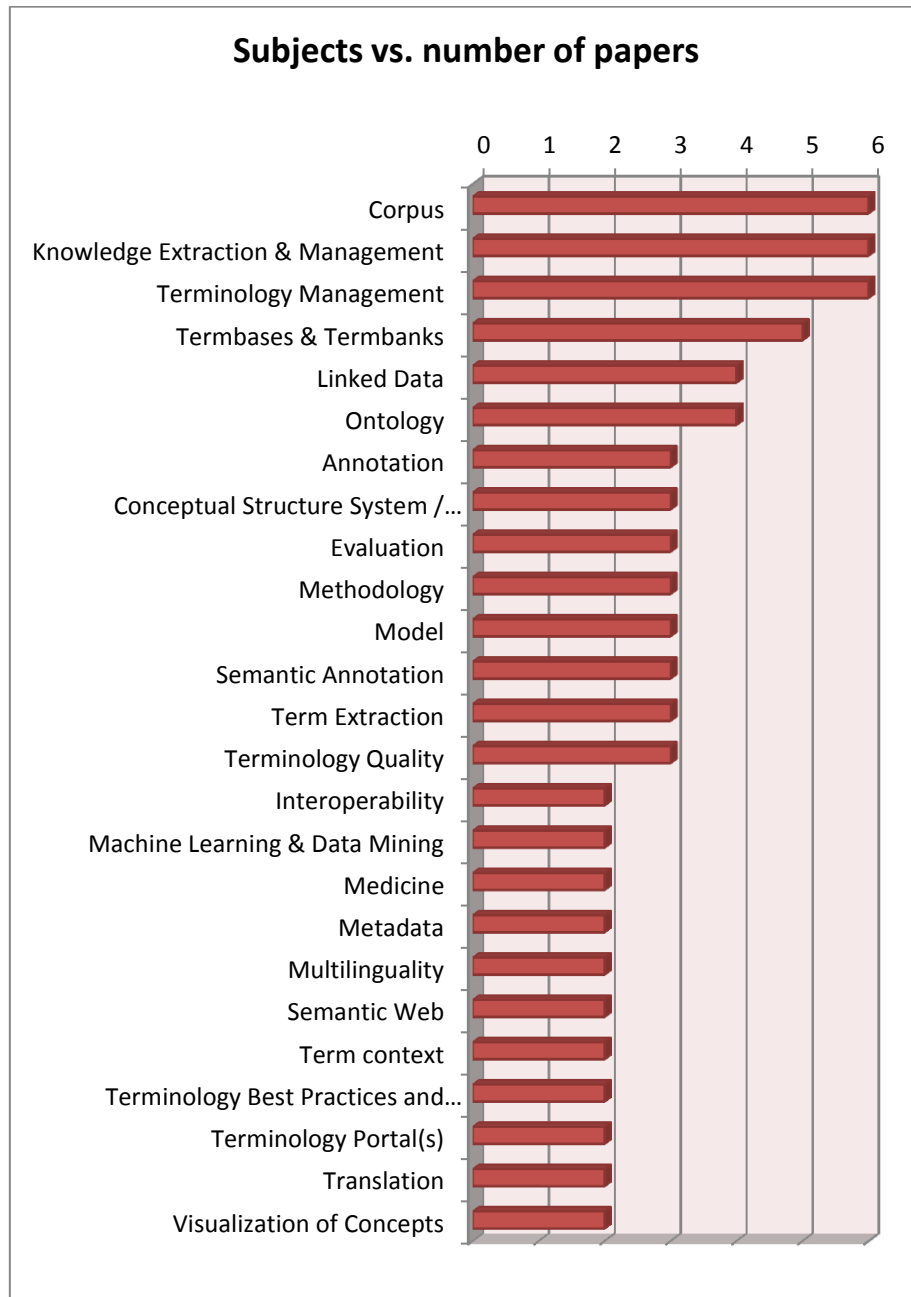
Thus, the papers presented at TKE 2016 fall within the following subjects: ‘Linked Open Data’, ‘Corpora for terminology work’, ‘Knowledge extraction’, ‘Knowledge organization’, ‘Term banks’ and ‘Terminology management’. A more detailed list of subjects and the number of papers addressing them is shown in the pie chart included in Figure 3.

Fig.2. Topics covered by TKE 2016 accepted papers.



Interestingly, however, only few papers related to the sub-theme ‘Linked Open Data’ were submitted, and therefore we are very happy that both keynote speakers, Eva M. Méndez Rodríguez and Sebastian Hellmann, will deal with this theme. The titles of the keynote speeches are ‘You call it Terminology, I call it Vocabularies: LOV for data in open science and cultural heritage’ (Eva M. Méndez Rodríguez) and ‘Challenges, Approaches and Future Work for Linguistic Linked Open Data (LLOD)’ (Sebastian Hellmann).

Fig. 3. Subjects addressed in TKE 2016 accepted papers, together with the number of papers addressing them.



Two workshops will take place in connection with TKE 2016.

At the workshop on *Terminology Teaching & Training*, issues in terminology teaching and training will be discussed in depth in so-called beehives, i.e. parallel controlled discussion group(s). This workshop is organized by:

- Henrik Nilsson (TNC, representing EAFT; Terminologicentrum TNC, Sweden)
- Christian Galinski (Infoterm, representing IITF; Infoterm, International Information Centre for Terminology)

The workshop *Making the visualization of concepts more attractive and smarter*, had a call for papers, and the 5 papers that were accepted are included in this volume. This second workshop is organized by:

- Professor Klaus Robering, University of Southern Denmark
- Associate Professor Lotte Weilgaard Christensen, University of Southern Denmark
- Associate Professor Rocio Chongtay, University of Southern Denmark
- Professor Bettina Berendt, KU Leuven, Belgium

We are grateful for the funding TKE 2016 received from the two Danish foundations, *Hedorfs Fond* and *Otto Mønstedts Fond*, and for the support from the Department of International Business Communication at Copenhagen Business School. We are also grateful to Emma Primdal Pedersen and Niklas Mellerup for the good job on editing this volume and handling other practical issues in connection with the conference. Finally, we would like to take the opportunity to explicitly thank Merete Borch for her continuous help with all kinds of issues concerning TKE 2016.

Bodil Nistrup Madsen
Antonio Pareja-Lora
Hanne Erdman Thomsen

Copenhagen, June 2016

Table of Contents

Termbanks & Terminology Management

BRUNO NAHOD

Can Big National Term Banks Maintain

Complex Cross-Domain Conceptual Relations?.....1

MIKI IWAI, KYO KAGEURA AND

KOICHI TAKEUCHI

Cross-lingual structural correspondence

between terminologies: The case of English and Japanese..... 14

HANNE ERDMAN THOMSEM

AND BODIL NISTRUP MADSEN

The DANTERM Model Revisited..... 24

Terminology, Quality and Evaluation

CRISTINA VALENTINI

Quality Control in Terminology Management..... 34

BARBARA HEINISCH-OBERMOSER

Web Interfaces of Terminological Databases that are

Available on the Internet from A Usability Perspective..... 44

FIRAS HMIDA, EMMANUEL MORIN,

BÉATRICE DAILLE AND EMMANUEL PLANAS

A Bilingual KRC Concordancer for Assisted Translation

Revision based on Specialized Comparable Corpora..... 54

Terminology, Corpora and Term Extraction

LOTTE WEILGAARD CHRISTENSEN

Semi-automatic Evaluation of

Terminological Web-crawled Corpora..... 64

GERHARD HEYER, CATHLEEN KANTNER, ANDREAS NIEKLER,

MAX OVERBECK AND GREGOR WIEDEMANN

Modeling the dynamics of domain

specific terminology in diachronic corpora..... 75

ALFREDO MALDONADO AND DAVE LEWIS

Self-tuning ongoing terminology extraction retrained on terminology validation decisions.....	91
--	-----------

Terminology and Semantic Annotations

ORNELLA WANDJI TCHAMI

Acquiring Vern Frames for a Text Simplification Lexicon in the Medical Domain.....	101
---	------------

MARÍA POZZI

Design of a corpus for mathematical knowledge transfer to 6-12 year old children.....	112
--	------------

ANTONIO PAREJA-LORA

Enabling Linked-Data-Based Semantic Annotations - the Ontological Modeling of Semantics in the OntoLingAnnot Model.....	124
--	------------

Terminology, Ontologies and Linked Data

JULIA BOSQUE-GIL, ELNA MONTIEL-PONSODA,
JORGE GRACIA AND GUADALUPE AGUADO-DE-CEA

Terminoteca RDF: a Gathering Point for Multilingual Terminologies in Spain.....	136
--	------------

BRUNO ALMEIDA, CHRISTOPHE ROCHE
AND RUTE COSTA

Terminology and ontology development in the domain of Islamic archaeology.....	147
---	------------

SARA CARVALHO, RUTE COSTA
AND CHRISTOPHE ROCHE

LESS Can Indeed BE More: Linguistic and Conceptual Challenges in the Age of Inoperability.....	157
---	------------

Terminology and Knowledge Organization Systems

ANDREAS LEDL AND JAKOB VOß

Describing Knowledge Organization Systems in BARTOC and JSKOS.....	168
---	------------

MARA ALAGIC, TATIANA OREL
AND GLYN RIMMINGTON

Toward Dynamic Representations of ThirdPlaceLearning.....	179
--	------------

TKE Workshop 2016:	
Making the Visualization of Concepts More Attractive and Smarter.....	188
ANITA NUOPPONEN	
Satellite System as a Visualization Tool for Concept Analysis.....	190
MARGARITA SORDO, CHRISTOPHER J. VITALE, PRIYARANJAN	
TOKACHICHU, DAN BOGATY, SAVERIO M. MAVIGLIA AND ROBERTO A. ROCHA	
Simple Graphical Representations of Ontology-based	
Clinical Decisions Support Knowledge Assets.....	201
BODIL NISTRUP MADSEN, SØREN BRIER, KATHERINE ELIZABETH LORENA	
JOHANSSON, BIRGER HJØRLAND, HANNE ERDMAN THOMSEN AND HENRIK SELSØE	
SØRENSEN	
The Landscape of Philosophy of Science.....	212
LOUISE PRAM NIELSEN	
Target Users' Diagrammatic Reasoning of	
Domain-specific Terminology.....	224
JESPER JENSEN AND LARS JOHNSEN	
Towards Concept Maps 3.0:	
Visual Learning Designs as Web Data.....	236

Can Big National Term Banks Maintain Complex Cross-Domain Conceptual Relations?

Bruno Nahod

Institute of Croatian Language and Linguistics, Croatia
{bnahod}@ihjj.hr

Abstract. This paper tries to answer the question ‘is the implementation of complex cross-domain conceptual relation possible in a multidomain term bank’. The Croatian term bank Struna, and the problems that terminologist working on it are facing will be used as a showcase. Some of the principles of, and proposed solutions for, the implementation of Domain Cognitive Models into the Croatian national term bank - Struna. A brief overview of the development of DCM will be given. Additionally, some of the major problems that occur when applying this type of sociocognitive-based conceptual structure to an existing objectivist (classical) hierarchical structure will be observed.

Keywords: terminology, Domain Cognitive Models, national term banks, Struna, conceptual structure

1 Introduction

In this paper we will present some of the principles of, and proposed solutions for, the implementation of Domain Cognitive Models (Nahod 2015b) into the Croatian national term bank - Struna. Additionally, we will try to identify some of the major problems that occur when applying this type of sociocognitive-based conceptual structure to an existing objectivist (classical) hierarchical structure.

2 Struna term bank

Struna is the Croatian National Term bank (<http://struna.ihjj.hr/>). Its aim is to gradually standardize Croatian terminology, for all professional domains, by coordinating the work of domain experts, terminologists and language experts. A broader objective of the program is to establish a framework for a national terminology terminological policy and to lay the foundations for more structured education in this field (Bratanić and Ostroški 2013b).

At the time of writing (early 2016) 18 domains have been processed and made public, with four additional domains in various stages of processing. Struna currently contains 31,256 concepts and close to 100,000 terms spanning ten languages: Croatian,

English, German, Italian, French, Latin, Russian, Slovenian, Czech and Slovakian. It must be noted that not all of the languages are equally represented.

Struna is primarily a normative terminological database. It is organized according to the relatively stable principles of the General Theory of Terminology (GTT) (Wüster 1979; Felber 1984) which is a, more or less, explicitly recognized approach to terminology planning, taking into consideration that ISO recommendations, for terminology management, are directly based on it. It was necessary to have a firm methodological framework for the practical terminographic work, because of the specific nature of the cooperation between field specialists at one end of the process and terminologists and other linguistic experts at the other end. In order to ensure that the various subject fields in the database are as structured and uniform in description, as they can possibly be, it is vital that a uniform approach is taken to any terminological description, guaranteed by a national terminology project environment. Conversely, the variety of domain knowledge included in Struna, as well as the various characteristics of each domain – conceptual structure and dynamics, specific communicative settings and intended users – meant modifications and adjustments were called for in the general terminological principles of the descriptions of particular domains (Bratanić and Ostroški 2013b, 667).

The workflow in Struna is organized into individual one-year projects. Each project is funded by the Croatian Science Foundation and each is, as far as coordination and the domain span is concerned, independent of the other Struna projects. When they propose projects, field experts define the domain that will be processed and all the terminological units that are edited in the project are automatically assigned to that domain. Terminologists from the Institute of Croatian Language and Linguistics have no influence over the process of choosing the projects or domains that will be processed.

This approach, consequently produced a steady inflow of problems, that were unsolvable in this strict GTT platform (Bergovec and Runjaić 2012), which, as a result, motivated researchers to seek more appropriate paradigms for terminology management (Bratanić and Ostroški 2013a).

2.1 The conceptual structure of Struna

The underlying conceptual structure of the Struna database must be simple. There are numerous reasons for this, the most important being the nature of the project's specifications, which were prescribed by the sole provider of funds, Croatian Science Foundation. The timeline for each project was 12 months, which meant there was an extremely short time-frame allowed for terminology processing. In any field of expertise, no matter how "small", and where field experts are expected to construct a conceptual structure, a robust and simple template is the only acceptable solution, given the time (Nahod 2009).

Conceptual structure featured in Struna's schema can be described as strictly vertical. The only relations explicitly defined are: the mandatory and semiautomatic affiliation of every concept into a specific field of expertise and a subordinate concept relation which is optional. "The Official Classification of Domains" (OCD) is used for defining domains, fields and branches (hr. područja, polja i grane). It is an official document, published by the Croatian government, which contains classifications of the scientific,

humanities and art domains. There is an implicit relation which can be understood from the definitions, given that most of them take the form of *genus proximum et differentia specifica*. In some special cases, additional information about conceptual relations is given in the notes, another optional data field, but such cases are actually so few in number that they can easily be omitted from any serious analysis.

As the number of domains (fields) and terminological units (concepts) grew in the Struna database, the shortcomings of this “robust” approach, to conceptual structure, began to show.

1. The Official classification of domains (OCD), that was used, was far from perfect and there were numerous cases where field experts could not agree on which field a concept belonged to, which led to multiple entries. This kind of problem is to be expected when one uses strict ‘is or is not’ rules of categorization. A partial solution, in the form of an interdisciplinary marker, was implemented. This marker was used to signify that a certain concept could conceivably belong to more than one field. However, this solution was flawed for the following two reasons:
 - a. The OCD already had an interdisciplinary field. This meant that when field experts were updating the classifications and users were looking at the interdisciplinary marker, they were often confused as to what the field label actually meant.
 - b. Field experts were reluctant to mark concepts as interdisciplinary, the main reason being that they felt that they were “giving away” their concepts to other domains.
2. As the number of terminological units grew, the search results became more and more confusing to the end users. Simple queries (one-word terms in the search engine’s simple mode) tended to produce multiple results that spanned manifold and often unrelated domains. These kinds of results effectively annulled even the simplest vertically based structure, because it began to appear that there was no structure (Figure 1).
3. A much deeper problem was also observed. A significant number of cases were found where a superordinate concept was defined in domains other than hyponyms. A combination of trying to restrict multiple entries and an inability to establish relations between two or more domains, had led to multiple discontinuities in the conceptual structure.

sila uzrok promjene gibanja čestice ili tijela	fizika
sila držanja kalupa sila koja održava pritanje dijelova kalupa tijekom ubrizgavanja i djelovanja naknadnoga tlaka u kalupu	strojarstvo
sila izboja poprečna sila koju stvara vijak i ostvaruje izboj krme	tehnologija prometa i transport
sila izlaska klipa sila koju stvara cilindar pri izlaznome hodu klipa	strojarstvo
sila izvlačenja sila koja djeluje na uzorak geosintetika pri izvlačenju iz uređaja za ispitivanje otpornosti na izvlačenje	građevinarstvo

Fig. 1. Partial results of a simple search for *sila* (en. force)

Figure 1 illustrates an example of the problems mentioned above. The first column contain concepts, as found by the search engine and the second column shows in which domain (project) the terminological unit was processed. Translations of the results and the domain affiliation as they appear on the list: *sila* – force / Physics, *sila držanja kalupa* – locking force / Machine engineering, *sila izboja* – transverse thrust force / Transportation technology, *sila izlaska klipa* – cylinder outstroke force / Machine engineering, *sila izvlačenja* – pullout force / Civil engineering. The first result listed is the concept “*sila*” (en. force) as defined in the field of physics; the results that follow it seem to be part of the same structure but are actually not connected to the first concept, in any formal way. This false representation of structured search results could conceivably lead the end-user into deploying the wrong term. In a hypothetical situation, an end user could click on a concept from transportation technology, while translating a text that is actually from the domain of physics, all the while believing he/she is still in the domain of physics. Thus, they could make an error in the target text. Many similar examples can be thought of where this situation might cause problems, or even mislead the end user.

2.2 Solution

It became evident that the existing structure would have to be upgraded, in order to solve the above-mentioned problems, plus some others. One of the first rules, when building the structure (scheme) of any database, is to design and develop it to a level that is as detailed as possible, before beginning to populate it with data. The reason for this is not that it is impossible to change something “on the go”, but rather that each change will entail further changes though every level of data, or in this case terminology management. So, any change in the structure of conceptual relations will inevitably require changes in the Content Management System (CMS), the search queries and algorithms, the Application Programming Interfaces (APIs), the search results sorts, the user manuals and even the web-page design, as a consequence.

Furthermore, there is the problem of the (re)editing the 30,000 + terminological units from the finished projects without the help of field experts or funding.

Finally, the Struna would most likely have to be closed to the public during software implementations or upgrades and during the editing of terminological units.

A solution that could bypass most, if not all, of these problems was proposed. It was to implement a conceptual structure as an overlay on the existing structure, without changing it.

3 Domain Cognitive Models

As a result of the effort to cope with the problems that emerged, on multiple levels, while processing terminological units in Struna, Domain Cognitive Models (DCM) (Nahod 2015b) emerged as the possible paradigm for processing specialized languages. DCM is based on Lakoff's idealized cognitive models (Lakoff 1987) and is highly influenced by sociocognitive linguistics and neuroscientific research.

As Nahod (sic.) has shown, enough evidence exists for us to regard field experts as a subgroup of society, or a subcultural community, that transcend definitional boundaries such as language, culture, and personal beliefs or preferences. The conceptual variations that can be observed in concepts, processed in Struna, can be better understood if they are described using cognitive models. The Domain-specific Cognitive Models were proposed for this purpose, to be schemata for concept clusters, in which certain specialized concepts form relatively stable relations and are categorized in different degrees of specialization in regards to their counterparts in other DCMs.

The DCM can be viewed as a collective conceptual subsystem, consisting of concepts that show variations in the properties that define them, in comparison to general or more broadly accepted conceptual systems. This variance emerges, either as a result of highly specialized (deeper) knowledge or as a result of research being focused on the specific properties (non-intrinsic) of the subject.

As more and more domains were processed in Struna, an interesting phenomenon was observed; "same" concepts were defined in different ways in different fields of expertise. As expected, it was observed that the more similar two domains were, the smaller the conceptual variations (differences in definition), and the more different the domains were the more pronounced the differences in definition. For example, the concept 'space' was defined in a similar way, and with almost exactly the same characteristics, in the domains of physics and mathematics, yet it was defined with a completely different set of properties in the domain of archaeology.

Maybe the most interesting finding was that not only do conceptual variations emerge between two domains that are essentially different (for example under the subjects of investigation and methodology etc. - i.e. physics and archaeology) but that the same conceptual variations exist when highly specialized subdomains (narrow field of interest) are compared with their main domains, for example crystallography and physics.

A hypothesis was proposed, that differences can be found in the conceptual structures, not only between two or more domains, but also between some domains and their subdomains. After a trial investigation using examples found in Struna, the initial hypothesis was expanded to include some of the observed phenomenon.

3.1 The layers in the conceptual structure

It was observed that certain clustered sets of concepts act as if they were independent of the general structure. A list of common properties was identified in most of the “sets”:

1. Small sets of related concepts often caused the biggest problems in harmonization procedures, because they invoked a perceivable semantic shift when compared to the main conceptual structure of a domain or even the whole of Struna’s structure.
2. Most of the sets concentrated around one concept which, more often than not, was presented in a one-word term.
3. All of the one-word terms, that were identified as focal concepts of these “independent sets”, were regular and high frequent words in the general Croatian language (Moguš, Bratanić, and Tadić 1999).
4. The sets were grouped together by the defining properties that were either considered unimportant, when making a definition on a more general level, or straight out wrong (Bergovec and Runjaić 2012; Nahod 2015b).

The most interesting finding was that these semantic clusters of concepts were not confined to the provided domain/project structure, but tended to go beyond that, finding their own niches in the conceptual structures of other domains. It became evident that this phenomenon was neither a local nor an insignificant one. Rather it was the result of a semantical environment, evoked by a highly specialized, or alternative strategy of categorization that was not anticipated when the Struna projects were planned.

As mentioned earlier, the domain structure in Struna can best be viewed as vertical and the whole structure of Struna itself can be viewed as a number of domains forming rows of vertical columns (Fig. 2), each populated by the terminological units processed in Struna projects, respectively.

Struna structure				
Physics	Mathematics	Archaeology	Chemistry	...

Fig. 2. A visualization of the domains in Struna

When we try to visualize these semantic clusters; those that can be identified through the conceptual relations that they form with concepts from other domains, appear as a layer of semantic shift that flows unceasingly across the “borders” of the domains. The concepts that produce this layer are not external to the structure of each domain, but are integral, although they are clearly recognized as being more strongly related to their subset than to the main conceptual structure of each domain.


Struna structure with DCM_1 layer				
Physics	Mathematics	Archaeology	Chemistry	...
				

Fig. 3. A rough visualization of a semantic layer in Struna

3.2 Theory behind praxis

The question was asked: do these concept clusters, or substructures, “misbehave” in Struna's structure, as well as in single domains, because the existing structure is inadequate or are the clusters just very different from the usual hierarchy-based structures. As stated earlier, the conceptual structure that underlies Struna is fairly simple and consequently quite robust, so it should be able to accommodate most, if not all, variations. That being said, the structure (of Struna) is mostly based on a conceptual structure, as represented by GTT (Felber 1984) and therefore is basically a product of classical, or objective, theory of categorization (Nahod and Vukša 2014). Considering how most modern researchers reject the “classical theory of categorization”, as a valid theory for explaining human cognition (Rosch 1978; Lakoff 1999; Murphy 2002; Gallese and Lakoff 2005; Roessler, Lerman, and Eilan 2011 *inter alia*)¹, this implies that Struna's structure is far from optimal. Conversely, there are no problems processing and structuring most of the concepts in Struna, so evidence does seem to suggest that the clustering variants, or semantic substructures, are behaving differently from the majority of concepts, for some reason.

The DCM paradigm emerged as an attempt to deal with the inconsistencies on the conceptual level of terminology management. As stated, the DCM is mostly based on the theoretical findings presented in idealized cognitive models. The main idea behind ICM is that each one structures a mental space, and when invoked it can cancel or change properties in the structure, outside of the ICM while at the same time retaining some of the other properties (Lakoff 1987, 68, 74).

If the same principles are applied to the “problematic” substructures of concepts described earlier, a valid model emerges to explain their behavior in a general conceptual structure, and possibly also the means to develop a model to describe them.

Therefore by using the same principles, we can attempt to explain the emergence of both the focal concepts of these clusters and the clusters themselves, as well as their ability to transcend their origin domains.

¹ The subject of human categorization and modern theories of it is too complex to include in this short paper. For further references please refer to overview of the subject by one of the eminent researchers i.e. G. Murphy (2002).

3.3 Experts as a subculture

The curtail presumption required for a cognitive model to be activated, is that a certain subgroup of people share specific knowledge that in some way varies from the knowledge that is shared with the general community. We strongly feel that, with the following observation, Sager et al (1980) are implicitly hinting that experts in a certain domain can be perceived as a highly affiliated subgroup.

“So we have to explain the fact that a French physicist can read an English research paper on physics or even understand a lecture on physics delivered in English while at the same time being incapable of reading an English newspaper or asking his way around London.” (Sager et al. 1980, 3)

If we assume that experts of a certain field behave as a sub-cultural community, then evoking a cognitive model can explain this phenomenon. This sub-culture evidently transcends the usual boundaries of nationality, language or even geographical distance. Following the same principle, there is no reason why a sub-subgroup, affiliated with a very narrow field of expertise, cannot transcend the more or less arbitrary domain bounds.

The main idea behind DCMs is to encircle the clusters of concepts that show semantic shift from the general categorization, in order to separate them from the main structure, while describing their connection to it and to other related domains (as defined in Struna). This will, presumably, result in a more accurate description of conceptual structures and consequently a better representation of terminological units. Furthermore, by implementing the layer based structure of DCMs into Struna’s schema, the means will be provided to present categorization variances to the end user.

4 On Implementation

Making changes to the structure of a database is never easy. As mentioned earlier, in a multi-domain term bank any change in the structure demands additional changes on all the levels of terminology management. As a way of avoiding making changes in the Struna’s schema that required taking it off-line for the duration their implementing, the DCM was envisioned as a complementary extension that could be superimposed over the existing structure (Nahod 2015b, 123).

Nahod (Nahod 2015a) describes the DCM extension in the XML form. Although the reported structure is still in development, the main elements correspond to the major properties of the DCM paradigm.

```

<domain id="01" name="fizika" altName="physics">
  <metData>
    <layers>
      <layer id="1_01" name="DCM_FF1" extendedness="classical"/>
      <layer id="11_01" name="DCM_FF11" extendedness="radiology"/>
      <layer id="2_01" name="DCM_FF2" extendedness="modern"/>
      <layer id="21_01" name="DCM_FF21" extendedness="relativistic"/>
      <layer id="22_01" name="DCM_FF22" extendedness="quantum"/>
      ...
    </layers>
  </metData>
  <CS altName="Conceptual Structure">
    <termUnit id="0001" layer="1_01">
    <termUnit id="0002" layer="2_01">
    <termUnit id="0003" layer="1_01">
    <termUnit id="0004" layer="21_01">
    <termUnit id="3210" layer="21_01">
    <termUnit id="3211" layer="21_01">
    <termUnit id="3212" layer="21_01">
    <termUnit id="3213" layer="21_01">
    <termUnit id="9801" layer="1_01">
    <termUnit id="1269" layer="11_01">
  </CS>
</domain>

```

Fig. 4. XML representation of the DCM structure (Source: Nahod 2015b, 115)

Figure 4 shows the part of the structure that contains the definition of the layers, as well as the identifying affiliation to the specific layer, for each terminological unit. As well as explicitly defining layers, the DCM extension also features the means to define the relations of different types between all concepts, regardless of their domain affiliation in the original Struna structure.

Figure 5 shows a part of the syntax, concerning conceptual relations in the DCM schema. This syntax allows for a broad spectrum of terms when defining conceptual relationships inside the same domain and/or the DCM, as well as across the whole database, regardless of the affiliation of either the source or the target concept. The example shown in Figure 5 is for concept **space** (three-dimensional), *prostor* in Croatian, declared in <term> tag. In the “relations” section, various relations can be seen. The relation to **n-dimensional space** (*n-dimenzijski prostor* in the Figure) illustrates the advanced way conceptual relationships can be stated: argument *type=“broader”* states the type of relationship, the *interlink* argument in the <target> element serves as a path to the DCM layer,; “2_01” - “modern physics” in Figure 4. And finally, the *term* argument states the term. By using this syntax, every terminological unit can be connected to any other in the whole Struna database, regardless of the type of conceptual relationship or the domain affiliation of the concepts. This feature is not possible in the default (current) Struna schema.

```

<CS altName="Conceptual Structure">
  <termUnit id="0001" layer="1_01">
    <langSet xml:lang='hr-HR'>
      <term type="preferredTerm">prostor</term>
      <term type="fullForm" >trodimenzijski prostor</term>
      <definition>neograničena trodimenzijska protežnost u kojoj tijela
        imaju relativne položaje</definition>
      <relations>
        <relation id="01" type="similar to">
          <target interLink="1_00" id="0001" lex="prostor" sense="1"/>
        </relation>
        <relation id="02" type="broader">
          <target interLink="2_01" id="0002" term="n-dimenzijski prostor"/>
        </relation>
        <relation id="03" type="hypernym">
          <target interLink="1_01" id="0003" term="euklidijski prostor"/>
        </relation>
      </relations>
    </langSet>
    <langSet xml:lang='en-EN'>
      <term>space</term>
    </langSet>
  </termUnit>

```

Fig. 5. XML representation of defining conceptual relations in DCM (Source: Nahod 2015b, 117)

As reported (Nahod 2015b), preliminary tests show promise and, considering that the simulation of search results seems to solve the problem of sorting results in Struna's current search engine (Figure 1), we do believe that further development of the model could result in an applicable extension to Struna.

By the end of 2016, a searchable alpha version of exemplary subset of terminological units (ca. 1000), with a DCM-based structure superimposed over them, will be open to the general public. Based on the testing, and on users' feedback, further research and development of the DCM will be proposed.

4.1 Challenges

There are a few obstacles to implementing DCM extension that can be noted. The biggest seems to be recognizing which concepts make up a certain layer. The original terminological units in Struna were entered into the database by field experts during each project's lifetime, which means that those experts are no longer available for further consultation².

The level of specialized knowledge that is required to identify the conceptual variations and the relations that cause semantic clusters or DCMs, far exceeds the knowledge that any terminologist currently working on Struna has.

Even assuming that the time and effort needed to develop this kind of knowledge level could be allocated; it is unlikely that anyone could do so for a "big" domain such as physics, let alone for the 18+ domains. This problem is further emphasized when we

² Except for a few highly motivated individuals that are, unfortunately, exemptions to the general consensus that the work on Struna is finished with the final project report.

consider how the nature of the DCM dictates, that the layers, defined in the database, must be able to traverse multiple domains and that it would be impossible to implement them properly, into just one domain.

It is evident that further consultation with field experts will have to be made mandatory, at that stage of implementation.

Furthermore, there is a question about how deep one should go when searching for semantic clusters. How important is it, for a multi-domain term bank, to have conceptual variations defined at the highest levels of expertise? This is a highly important question considering that the end users of Struna show such a wide spectrum of profiles: from high school pupils to field experts and from the general public to specialized translators.

It is our strong opinion that even should the funds be available, the cost of completely recoding all of the terminological units in Struna, according to schema based on the DCMs would, more than likely, exceed the benefits.

Therefore it will be necessary to adapt the DCM model, into a more practical approach, that would target only the most problematic cases, from the users' point of view, as well develop some kind of automatization, preferably a corpus based one.

5 Conclusion

Before attempting to answer the question posed in the title of this paper, we should consider the implications of observations that have been made about all big multi-domain terminological databases and not just Struna.

A good example, of working cognitive-based terminology processing, is the Eco-Lexicon with a Frame based terminology model (Faber et al. 2006; Faber and Castro 2014). It is based on the frame semantic (Fillmore 1985) that the LexiCon Research Group has been developing, for the last 13 years, and is a highly detailed conceptual structure in the domain of Environment. To our knowledge, it is, currently, the only fully functional cognitive-based terminological database. Although it is much more detailed than DCM, and is only applied to one specialized domain, it is still a good reference point when considering how much effort and funding will be required to develop such a complex conceptual structure for terminology management. If we take the Eco-Lexicon timeframe as a reference and try to calculate how much time it would take to implement even a simpler structure, as assumed by DCM, into a relatively small database such as Struna, we soon come to the incomprehensible number of 100+ years.

To return to the paper's main question: implementing a complex conceptual structure over a multi-domain terminological database is not something that can be done easily, and is probably not even possible to do it *ad-hoc* for the big database as Struna. It would, and will be, essential to develop a means to cluster and code the concepts (semi)automatically.

In the first stage of implementing DCMs into Struna, a manual coding will be applied, but only on exemplary cases – that is on the most problematic conceptual variations that have been observed so far. After the testing period we expect to implement a crucial modification into our model that will determine any further development.

Acknowledgements

Research reported in this paper was co-financed by the European Social Fund (Project HR.3.2.01-0072) within OP Human Resources Development 2007–2013.

References

- Bergovec, Marina, and Siniša Runjaić. 2012. "Harmonization of Multiple Entries in the Terminology Database Struna (Croatian Special Field Terminology) 231–241." In *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012)*, edited by Guadalupe Aguado de Cea and Al., 231–41. Madrid.
- Bratanić, Maja, and Ana Ostroški. 2013a. "STRUNA: National Croatian LSP Term Base Creation – Challenges and Lessons Learned." In *Specialised Lexicography. Print and Digital, Specialised Dictionaries, Databases*, edited by Vida Jesenšek, 83–93. Berlin/Boston: De Gruyter.
- . 2013b. "The Croatian National Termbank STRUNA: A New Platform for Terminological Work." *Collegium Antropologicum* 37 (3): 677–83.
- Faber, Pamela, and Miriam Buendía Castro. 2014. "EcoLexicon." In *XVI EURALEX International Congress: The User in Focus*, edited by Andrea Abel, Chiara Vettori, and Natascia Ralli, 601–6. Bolzano: Institute for Specialised Communication and Multilingualism.
- Faber, Pamela, Silvia Montero Martínez, María Rosa Castro Prieto, José Senso Ruiz, Juan Antonio Prieto Velasco, Pilar León Araúz, Carlos Márquez Linares, and Miguel Vega Expósito. 2006. "Process-Oriented Terminology Management in the Domain of Coastal Engineering." *Terminology* 12 (2): 189–213. doi:10.1075/term.12.2.03fab.
- Felber, Helmut. 1984. *Terminology Manual*. Vienna: Infoterm.
- Fillmore, Charles J. 1985. "Frames and the Semantics of Understanding." *Quaderni Di Semantica* 6 (2): 222–54.
- Gallese, Vittorio, and George Lakoff. 2005. "The Brain's Concepts: The Role of the Sensory-Motor System in Conceptual Knowledge." *Cognitive Neuropsychology* 22 (3): 455–79. doi:10.1080/02643290442000310.
- Lakoff, George. 1987. *Women, Fire, and Dangerous Things. Artificial Intelligence*. Vol. 35. University of Chicago Press Chicago. doi:10.1016/0004-3702(88)90035-5.
- . 1999. "Cognitive Models and Prototype Theory." In *Concepts: Core Readings*, edited by Eric Margolis and Stephen Laurence, 391–422. MIT Press.
- Moguš, Milan, Maja Bratanić, and Marko Tadić. 1999. *Hrvatski čestotni rječnik*. Zagreb: Zavod za lingvistiku Filozofskog fakulteta Sveučilišta u Zagrebu, Školska knjiga.
- Murphy, Gregory L. 2002. *The Big Book of Concepts*. Cambridge, MA: MIT Press.
- Nahod, Bruno. 2009. "Baza Podataka." In *Hrvatski terminološki priručnik*, edited by Dunja Brozović Rončević, 101–4.
- . 2015a. "Brak čestice i prostora: Sociokognitivna Poredbena Analiza

- Pojmovnih Struktura Strukovnih Jezika Fizike I Antropologije.” In *Od Šuleka do Schengena: terminološki, terminografski i prijevodni aspekti jezika struke*, edited by Maja Bratanić, Ivana Brač, and Prichard Boris, 169–96. Zagreb: Institut za hrvatski jezik i jezikoslovlje.
- . 2015b. “Domain – Specific Cognitive Models in a Multi – Domain Term Base.” *Suvremena lingvistika* 41 (80): 105–28.
- Nahod, Bruno, and Perina Vukša. 2014. “On Problems in Defining Abstract and Metaphysical Concepts – Emergence of a New Model.” *Collegium Antropologicum* 38 (Sup. 2): 181–90.
- Perception, Causation, and Objectivity*. 2011. Oxford: Oxford University Press.
- Rosch, Eleanor. 1978. “Principles of Categorization.” *Cognition and Categorization*, 27–48.
- Sager, Juan C., David Dungworth, and Peter F. McDonald. 1980. *English Special Languages. Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter.
- Wüster, E. 1979. *Einführung in Die Allgemeine Terminologielehre Und Terminologische Lexikographie*. Wiena/New York: Springer.

Cross-lingual structural correspondence between terminologies: The case of English and Japanese

Miki Iwai¹, Koichi Takeuchi², and Kyo Kageura³

¹ Graduate School of Interdisciplinary Information Studies, The University of Tokyo

² Graduate School of Natural Science and Technology, Okayama University

³ Graduate School of Education, The University of Tokyo

¹ 1156553643@mail.ecc.u-tokyo.ac.jp

² koichi@cl.cs.okayama-u.ac.jp

³ kyo@p.u-tokyo.ac.jp

Abstract. This paper analyses the structural correspondence between English and Japanese terminologies. Terminologies contain many complex terms, and each constituent element of terms represents an important conceptual feature. We investigated the structural correspondence by (a) constructing a terminology network for English and Japanese separately, (b) identifying structural characteristics of the network by decomposing the network into components, and (c) analysing the degree of correspondence among components in English and Japanese. We used terminologies of two domains, i.e. computer science and economics.

Keywords: terminology structure, cross-lingual correspondence, network analysis

1 Introduction

This paper analyses the structural correspondence between English and Japanese terminologies.

Terminologies in most languages contain a substantial number of complex terms, irrespective of domain (Cerbah, 2000; Nomura and Ishii, 1989). This reflects the fact that terminologies tend towards systematically representing concepts (Kageura, 2012, 2015), i.e. each constituent element of a term represents an important feature of the concept represented by the term, and terms that represent related concepts tend to show their relationship through common constituent elements⁴ (Sager, 1991).

Work in bilingual term extraction often adopts the assumption that there is substantial cross-lingual correspondence between complex terms, as indicated by the use of the “compositional translation” approach (Delpech and et al., 2012; Morin and Daille, 2010, 2012; Tonoike and et al., 2005) and the “bilingual extrapolation” approach (Sato and et al., 2013).

⁴ This, incidentally, is what Saussure (de Saussure, 1910-11) called “relative motivation”.

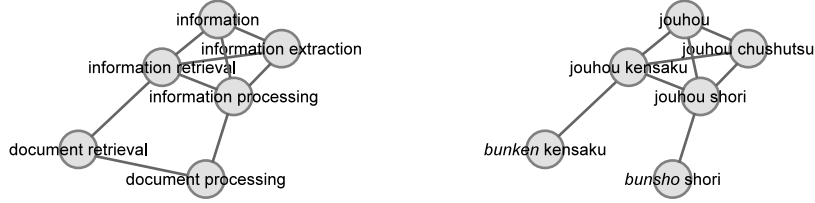


Fig. 1. Terminology networks of the putative terminology in English and Japanese

However, there is little work that empirically investigates the cross-lingual correspondence of the structure of terminologies across languages (Asaishi and Kageura, 2011). We explored this issue by (a) constructing terminology networks for each language using bilingual terminologies, (b) applying partitive clustering algorithms to the terminologies of each language, and then (c) analysing the overlap and difference between the term clusters generated in the two languages.

The paper is organized as follows. In section 2, we first elaborate on what we observed in this study and then introduce the clustering algorithms or community detection algorithms we used in this study. In section 3, we detail the data and the experimental setup. In section 4, we examine and analyse systematicity and correspondence between English and Japanese terminologies.

2 Terminology network and clustering

2.1 Cross-lingual mapping of terminological structure

Systematicity of a terminology can be grasped approximately by defining the terminology network – in which terms constitute vertices and common constituent elements constitute edges – and by analysing its characteristics (Kageura, 2012). Suppose we have a terminology consisting of six terms, i.e. “information”, “information retrieval”, “information extraction”, “document retrieval”, “document processing”, and “information processing”. Figure 1 (left) shows the terminological network constructed from this terminology. Naturally,

$$\begin{aligned} \text{degree}(v_i) &\simeq \sum_j \text{frequency}(c_{ij}) \\ \text{weight}(e_{ik}) &= |\{c_{ij}\} \cap \{c_{kl}\}| \end{aligned}$$

where v_i is the vertex (term) with index i , c_{ij} is a j -th constituent of the term v_i , and e_{ik} , which is defined by the number of common constituent elements between the two terms v_i and v_k , is the edge between v_i and v_k .

As terms are located between artificial nomenclature and ordinary words (Kageura, 2015), the nature of systematicity of a terminology differs from one language to another. In the above example, for instance, the corresponding Japanese terms

are “jouhou”, “jouhou kensaku”, “jouhou chushutsu”, “*bunken* kensaku”, “*bunsho* shori”, and “jouhou shori” – English “document” corresponds to “bunken” and “bunsho”. The Japanese terms have a different degree of systematicity (Figure 1, right).

To observe the nature of systematicity of terminologies and to analyse the cross-lingual structural correspondence of terminologies, we applied partitive clustering or community detection to the terminological network and observed the nature of clusters or components. This approach can also be used for generating bilingual potential term candidates for automatic bilingual terminology augmentation (Sato and et al., 2013).

2.2 Community detection or clustering algorithms

Many methods have been proposed to divide a graph into clusters (Clauset and et al., 2004; Rosvall and Bergstrom, 2008; Raghavan and et al., 2007; Blondel and et al., 2008; Pons and Latapy, 2006; Newman, 2006). After we examined the main methods, we decided to use the Potts spinglass-based approach (Reichardt and Bornholdt, 2006), as it is held to produce good results (Orman and Labatut, 2009).

The Potts spin glass model is a heuristic algorithm for solving the global optimization problem (Kirkpatrick, 1984). This method has the advantage of being able to obtain the globally optimal solution.

Formally, the Potts model consists of a lattice of N sites, on each of which is placed a spin that can take q -states. The Hamiltonian H is given as:

$$H(\{s\}) = - \sum_{(i,j)} J_{ij} \delta(s_i, s_j)$$

where J_{ij} represents the strength of relationship between s_i and s_j , and s_i and s_j represent the state of spins i and j . δ returns 1 when s_i and s_j are in an identical state and 0 otherwise. For the community detection or clustering of a network, Reichardt and Bornholdt (2006) proposed the Hamiltonian or cost function to be minimized as:

$$H(\{\sigma\}) = - \sum_{i \neq j} (W_{ij} - \gamma p_{ij}) \delta(\sigma_i, \sigma_j)$$

where W_{ij} denotes the adjacency matrix of the graph with normalized weight for edges, p_{ij} denotes the null model that gives the baseline probability that a link exists between node i and j , and γ is the parameter for distributing the weight between the reward for internal links and the penalty for internal nonlinks. This Hamiltonian can be transformed to the modularity Q provided in equation (4) above, in which higher Q corresponds to lower H .

Table 1. The distribution of terms in each terminology

Dom.	Lang.	T	1	2	3	4+
Com.	En	16259	2634(16.20%)	9044(55.62%)	3645(22.42%)	936(5.76%)
	Ja	16259	2002(12.31%)	7141(43.92%)	4782(29.41%)	2334(14.36%)
Ecn.	En	9120	1219(13.37%)	4858(53.27%)	1659(18.19%)	1384(15.17%)
	Ja	9120	947(10.38%)	3753(41.15%)	2814(30.86%)	1606(17.61%)

Table 2. Basic quantities of terminologies and terminological networks

Dom.	Lang.	T	N	fw	V	E	S
Com.	En	16259	5563	510	14186	992319	1100
	Ja	16259	4803	6634	15062	998245	1468
Ecn.	En	9120	5420	1958	8922	278836	749
	Ja	9120	4647	4691	9119	267603	863

3 Data preparation and setup

3.1 Terminology data and pre-processing

We used Japanese-English bilingual terminologies in the field of computer science (Aiso, 1993) and in the field of economics (Yuhikaku, 1986). Table 1 shows the number and ratio of terms by length in each terminology, i.e. single terms, terms with two constituents, terms with three constituents and terms with four or more constituents. In table 1, “Dom” stands for the domain, “Lang” the language, and the T the number of terms. It can be observed that two-word terms are dominant.

For Japanese terms, we (a) first divided them into constituent elements using MeCab⁵; (b) removed functional elements; and (c) removed dependent verbs and auxiliaries. For MeCab, we used UniDic⁶, which is designed to consistently identify the smallest meaningful units for Japanese. For English terms, we (a) lemmatised constituent words of each term using a lemmatiser⁷; and (b) removed function words.

After the pre-processing, we generated the network for English and Japanese terminologies independently. We used python and python igraph library for network analysis⁸. Table 2 shows the basic quantities of the data and the generated network. In Table 2, N is the number of constituent elements and fw the number of functional words. V and E stand for the numbers of vertices and of edges, and S the number of isolated terms. Among the most frequent constituents in computer science are “data”, “processing”, and “system” for English and “システム (system)”, “データ (data)”, and “通信 (communication)” for Japanese. In economics, they are “insurance”, “rate”, and “system” for English and “資本

⁵ <http://taku910.github.io/mecab/>

⁶ http://pj.ninjal.ac.jp/corpus_center/unidic/

⁷ <http://www.nltk.org/api/nltk.stem.html>

⁸ <http://igraph.org/>

Table 3. Basic data of components of terminology networks

Dom.	Lang.	#cmp.	max subgraph			second subgraph	
			V	E	D	V	E
Com.	En	1118	13046	992293	0.0117	3	2/3
	Ja	1544	13380	998034	0.0112	8	12/23
Ecn.	En	772	8127	278812	0.0084	3	2
	Ja	929	8096	267484	0.0082	6	7

Table 4. The results of clustering

Dom.	Lang.	#cmp	max	min	sdev.	#vertices of clusters
Com.	En	24	1571	11	400.18	[1571, 1229, 1094, 846, 842, 827, 759, 713, 691, 673, 587, 528, 524, 511, 464, 286, 285, 265, 109, 103, 56, 49, 23, 11]
	Ja	24	1491	25	376.00	[1491, 1217, 1170, 1055, 850, 839, 707, 692, 597, 498, 488, 484, 475, 467, 465, 399, 382, 258, 246, 202, 166, 133, 74, 25]
Ecn.	En	25	789	17	184.60	[789, 645, 544, 495, 484, 483, 447, 422, 409, 402, 377, 333, 298, 281, 266, 225, 215, 183, 180, 174, 139, 130, 118, 71, 17]
	Ja	25	675	28	175.54	[675, 657, 612, 470, 469, 466, 447, 433, 374, 374, 372, 357, 342, 336, 231, 212, 209, 206, 199, 172, 145, 136, 103, 71, 28]
Com.	En	10	2398	525	624.00	[2398, 1918, 1673, 1672, 1511, 1460, 678, 663, 548, 525]
	Ja	10	2290	586	620.39	[2290, 2289, 1696, 1688, 1531, 1082, 896, 721, 601, 586]
Ecn.	En	10	1513	373	293.59	[1513, 1120, 890, 751, 733, 717, 706, 691, 633, 373]
	Ja	10	1389	346	273.12	[1389, 948, 927, 908, 847, 840, 744, 705, 442, 346]

(capital)”, “経済 (economy)”, and “保険 (insurance)” for Japanese. We can observe differences between English and Japanese terminologies. Among isolated terms (S in Table 2), 110 are common between the English and Japanese in computer science, while 104 are common in economics.

3.2 Largest components

Each of the terminology networks consists of a single giant component (max subgraph) and many smaller components including isolates. Table 3 shows the data of components, in which “#cmp” indicates the number of components, V the number of vertices, E the number of edges and D the density (only given for max subgraph). As the max subgraphs attract most of the terms and constitute the core part of terminological structure, we analysed the cross-lingual structural correspondence of the max subgraphs by applying the algorithm based on the Potts spin glass model and extracting clusters. Incidentally, in the terminology of computer science, English and Japanese max subgraphs share 11533 common (corresponding) terms (88.4% for English and 86.2% for Japanese). In the domain of economics, they are 7241 (89.1% for English and 98.4% for Japanese).

4 Cross-lingual correspondence of clusters

We created clusters from max subgraphs, setting the number of spins (clusters) as 25 and 10⁹. Table 4 shows the basic quantities of clusters, where max and min show the number of vertices in the largest and smallest clusters.

⁹ This was heuristically decided by referring to subdomains of the fields.

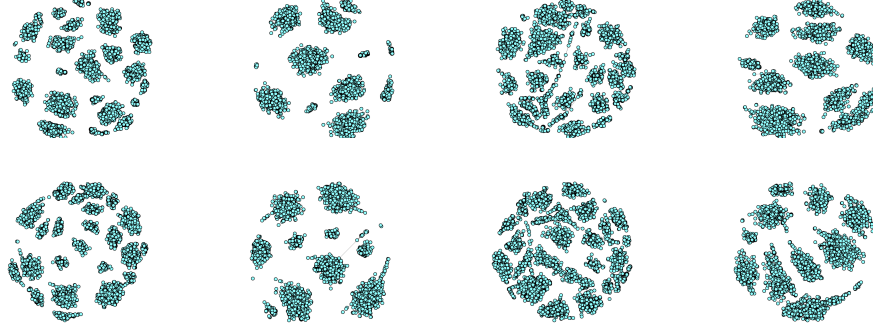


Fig. 2. Cluster visualization (top: English; bottom: Japanese; left to right: computer science 25 and 10, economics 25 and 10)

Table 5. Density and diameter of clusters

Dom.	Lang.	#cmp	Density				Diameter			
			max	min	mean	sdev	max	min	mean	sdev
Com	En	24	0.90	0.03	0.29	0.26	9	1	6.17	1.88
	Ja	24	0.66	0.04	0.20	0.17	9	1	6.17	1.88
Ecn	En	25	0.55	0.02	0.18	0.13	11	3	7.24	1.76
	Ja	25	0.61	0.04	0.18	0.16	11	3	7.84	2.01
Com	En	10	0.89	0.01	0.22	0.30	9	3	7.2	1.93
	Ja	10	0.50	0.02	0.19	0.19	11	4	7.6	2.22
Ecn	En	10	0.36	0.02	0.10	0.10	11	6	8.4	1.35
	Ja	10	0.56	0.02	0.12	0.17	14	6	9.3	2.36

4.1 Formal characteristics of clusters

Let us first observe the formal characteristics of the clusters, without delving into the cross-lingual correspondences based on bilingual term pairings. Figure 2 is a visualization of the clusters of the max subgraph. It indicates that, at least to the eye, clusters for English and Japanese terminologies of the same domain show much higher similarity than the clusters of computer science and economics terminologies within one language. This can be confirmed by the distribution of the number of vertices given in Table 4; for both 25 (or 24) clusters and 10 clusters, patterns of distribution of the number of vertices as observed from the largest cluster, smallest cluster and the standard deviation (sdev) are similar between English and Japanese in the same domain.

To be more analytically rigid, we observed the distribution of density and diameter of the clusters. Table 5 gives the maximum, minimum, mean and standard deviation of the density and diameter of the clusters. Density shows complex patterns, i.e. mean, standard deviation and maximum density indicate that there is some similarity between English and Japanese in the same domain, although some values (e.g. minimum density) seem to reflect more the language-dependent characteristics. The values of diameter show, on the other hand, similar tendencies between English and Japanese terminologies of the same domain. All in all,

Analysis of the structural correspondence between different languages

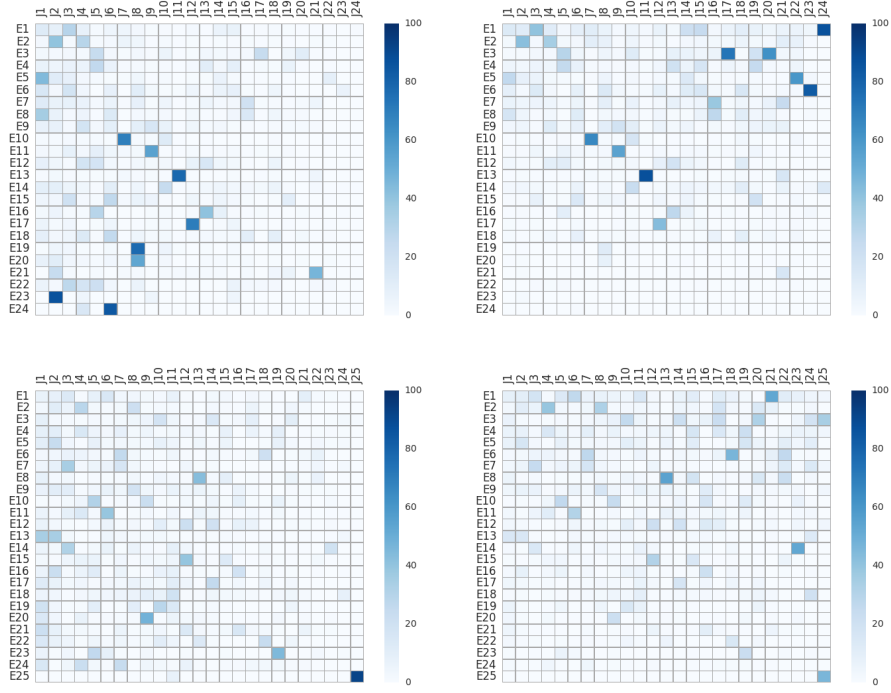


Fig. 3. Correspondence of terms in each of the 25 clusters (top: computer science; bottom: economics; left to right: the percentage of correspondence for each term based on English terms and Japanese terms)

we can reasonably conclude that there is a certain degree of formal cross-lingual correspondence within the corresponding terminology of the same domain.

4.2 Term-level cross-lingual correspondence of clusters

Figure 3 and Figure 4 show the ratio of corresponding terms (vertices) in each cluster between English and Japanese terminologies. The darker the cell color, the higher the ratio of corresponding terms. The panels on the left show the ratio of corresponding terms for English clusters, while those on the right show the ratio for Japanese clusters. Henceforth, we will describe that the overlap between English and Japanese clusters as “strong” if 40% or more of the terms in a cluster in one language belong to a single cluster in the other language, “very strong” if the overlap is 60% or more, “reasonable” if the overlap is between 20 and 40%, and “weak” if the overlap is 10 to 20%. For succinctness, we use “com25”, “ecn10” etc. as an abbreviation for the domain plus number of clusters. E1, J1, etc. refer to cluster id for English and for Japanese.

For com25, we can see that E2 and J2, E10 and J7, E11 and J9, E13 and J11, E17 and J12 have strong *mutual* overlap of terms, among which E10 and J7 and

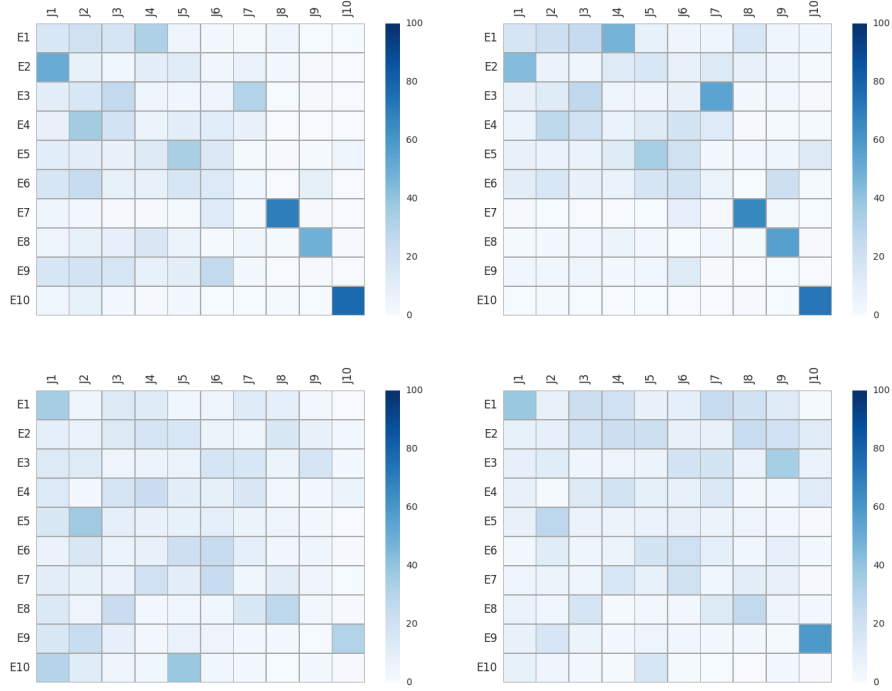


Fig. 4. Correspondence of terms in each of the 10 clusters (top: computer science; bottom: economics; left to right: the percentage of correspondence for each term based on English clusters and Japanese clusters)

E13 and J11 have very strong mutual overlap. The size of clusters with strong mutual overlap resides in the middle range except for E2 and J2. Larger clusters tend to have one to several reasonable and weak overlap, as shown between E1 and J1/J3, J1 and E1/E5/E8, E3 and J5/J17/J20. If we compare the panel on the left and the panel on the right for com25 (Figure 3), we can observe a symmetric relation. This implies that smaller clusters in one language constitute part of larger clusters in the other.

For ecn25, we can see that E8 and J13 and E25 and J25 have strong mutual overlap, while several clusters have reasonable mutual overlap, as shown by E2 and J4/J8, E6 and J7/J18, E10 and J5/J9, E11 and J6, E12 and J12, and E15 and J12. In ecn25, the number of cluster with strong or very strong mutual overlap is smaller than com25. A symmetric relationship can be observed also in ecn25, but less saliently than in com25.

We can interpret the patterns of overlap in com10 and ecn10 in a way consistent with the patterns observed in com25 and ecn25. For com10, we have very strong mutual overlap between E7 and J8 and E10 and J10, and strong mutual overlap between E2 and J1 and E8 and J9. Except for E2 and J1, the

clusters with strong mutual overlaps are smaller clusters (about the same size as the middle-sized clusters in com25). Larger clusters tend to have a reasonable degree of mutual overlap. The symmetric pattern is much less salient in com10 than in com25, although vague tendencies can be identified.

For ecn10, there is no strong mutual overlap. There are six reasonable mutual overlaps, i.e. E1 and J1, E5 and J2, E6 and J6, E7 and J6, E8 and J8, and E9 and J10.

General tendencies of cross-lingual correspondences can be summarised as follows:

- the degree of cross-lingual correspondence between English and Japanese is higher in computer science than in economics;
- middle-sized clusters tend to have stronger mutual overlap, while the larger clusters tend to have several to several overlaps;
- when the number of divisions is large, smaller clusters in one language tend to constitute part of larger clusters in the other language.

5 Conclusion and outlook

We observed structural correspondence between English and Japanese terminologies of computer science and economics. The analysis has revealed that there is a different degree of correspondence both in the *form* of structure and in the degree of term-level matching of the structure, depending on the domain. Theoretically, the present work sheds light on the nature of the systematicity of terminologies from the cross-lingual point of view, while at the same time showing the usefulness of the approach for theoretical analysis. We plan to further elaborate on the tendencies observed in this study, taking into account such factors as terminology size, clustering algorithm and cluster size. We will also extend the data to terminologies of other domains and other language combinations.

From the application point of view, that bilingual correspondence of terms between monolingually created clusters is not necessarily high implies that the “compositional translation” and “bilingual extrapolation” approaches may create different results if source and target languages are swapped and thus bidirectional treatment will have merit for dealing with bilingual terminologies (Sato and et al., 2013). We will explore the potential of bidirectional potential term extrapolation for term crawling.

Acknowledgments

This study is supported by JSPS Grant-in-Aid (A) 25240051 “Archiving and using translation knowledge to construct collaborative translation training aid system.”

References

H. Aiso. *Dictionary of Information Processing*. Tokyo: Ohm, 1993.

- T. Asaishi and K. Kageura. Comparative analysis of the motivatedness structure of Japanese and English terminologies. *In Proc. 9th TAI*, pages 38–44, 2011.
- V. D. Blondel and et al. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 2008.
- F. Cerbah. Exogeneous and endogeneous approaches to semantic categorization of unknown technical terms. *In Proc. 18th COLING*, pages 145–151, 2000.
- A. Clauset and et al. Finding Community Structure in Very Large Networks. *Physical Review*, 70:66–111, 2004.
- F. de Saussure. *Troisième Cours de Linguistique Générale - noté par Émile Constantin*. Geneva: Bibliotheque publique et universitaire, 1910–11.
- E. Delpéch and et al. Extraction of domain-specific bilingual lexicon from comparable corpora: Compositional translation and ranking. *In Proc. 24th COLING*, pages 745–762, 2012.
- K. Kageura. *The Quantitative Analysis of the Structure and Dynamics of Terminologies*. Amsterdam: John Benjamins, 2012.
- K. Kageura. Terminology and lexicography. *Handbook of Terminology*, 1:45–59, 2015.
- S. Kirkpatrick. Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics*, pages 975–986, 1984.
- E. Morin and B. Daille. Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation*, 44:79–95, 2010.
- E. Morin and B. Daille. Revising the compositional method for terminology acquisition from comparable corpora. *In Proc. 24th COLING*, pages 1797–1810, 2012.
- M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* 74, 2006.
- M. Nomura and M. Ishii. List of Stems in Japanese Technical Terms. Technical report, National Language Institute, Tokyo, 1989.
- G. Orman and V. Labatut. A comparison of community detection algorithms on artificial networks. *In Discovery Science*, pages 242–256, 2009.
- P. Pons and M. Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2):191–218, 2006.
- U. N. Raghavan and et al. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 2007.
- J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E* 74, 2006.
- M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *In Proc. PNAS*, 105(4):1118–1123, 2008.
- J. C. Sager. *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins, 1991.
- K. Sato and et al. Terminology-driven Augmentation of Bilingual Terminologies. *In Proc. MT summit XIV*, pages 3–10, 9 2013.
- M. Tonoike and et al. Effect of domain-specific corpus in compositional translation estimation for technical terms. *In Proc. 2nd IJCNLP*, pages 116–121, 2005.
- Yuhikaku. *Dictionary of Economy Terms*. Yuhikaku, Tokyo, 1986.

The DANTERM Model Revisited

Bodil Nistrup Madsen and Hanne Erdman Thomsen

Copenhagen Business School, Copenhagen, Denmark
{bnm,het}.ibc@cbs.dk

Abstract. In this paper we present the history of a model for a terminology management system, be it for big national term banks or for corporate term-bases. We go through the development stages of the DANTERM Model, starting in the late 1970's where the main purpose was translation, and advanced IT systems were not available, and ending today where more advanced terminology management systems (TMS) exist. Furthermore, systems for automatic extraction of information about concepts, automatic construction and consistency checking of concept systems (terminological ontologies) are now being developed. Both advanced TMS systems and automatic systems require new terminological data categories. Despite the existence of advanced TMS's, a very important problem has not been solved, namely the proper handling of equivalence relationships in multicultural terminology. We propose a solution to this problem and encourage ISO TC 37 committees, developing standards which prescribe data structures for terminological data, to incorporate the proposed solution.

Keywords. Terminology management systems, Data categories, Terminological data modeling, Terminological ontologies.

1 Introduction

In Denmark there is a long tradition for terminology work. Researchers from the Copenhagen Business School (CBS) and the Southern Danish University (SDU), two strong terminology research environments, have co-operated for many years, among others in a large research project, supported by the Danish Research Council. The researchers of these two institutions have developed principles and tools for terminology work, including the first Danish model for term bank entries, the DANTERM Model.

Terminology databases are important for translation purposes. Even so, it has been difficult in recent years to get funding for multilingual terminology work and research in Denmark.

However, during the last 10 years, Danish public authorities have been very interested in monolingual (i.e. Danish) concept clarification as a basis for clear communication with citizens (e.g. in self-service systems), for the development of IT systems and for data exchange. Therefore they have encouraged the development of new

methods and tools for the handling of information about concepts in concept systems or terminological ontologies.

Below we will report on the results of the research on and development of the model for terminology management systems carried out at CBS.

2 Once upon a time: the early model

The DANTERM Model was originally developed by the Terminology Centre at the Copenhagen Business School in the late 1970's and the early 1980's. It comprised a comprehensive set of data categories and a structure for a Danish Terminology Bank, which could meet the needs of all Danish users. The model was implemented in a central database system (Engel & Madsen, 1985), the results of students' and teachers' terminology work were stored, but unfortunately it was not possible to obtain resources for large-scale production of terminological data or for service functions. The full model is described in The DANLEX project group (1979) and The DANTERM project group (1987).

In parallel with the development of the international standard on data categories for terminology management, ISO 12620:1999, the Danish Standardization Organization developed a standard comprising a taxonomy for the classification of Lexical data categories, STANLEX (DS 2394-1, 1998). This was mainly because there was a need for a standard covering not only terminological data categories, but also data categories used in lexicographical data collections and in lexica of software for natural language processing. Consequently there was also a need for a systematic structure that was able to cover all these kinds of data collections. In STANLEX the main groups of information types are structured according to the main linguistic disciplines: etymological information, grammatical information, graphical information, phonetic information, semantic information and usage. The data categories in The DANTERM Model gave input to the work on this standard.

Appendix 1 presents a subset of the DANTERM categories illustrating the overall structure of the DANTERM entry and main principles, namely that

- all concept-related data categories are repeated for each language,
- all term-related data categories are repeated for each term.

Furthermore each entry contains language-independent information which appears only once. In addition to terminological entries, the model also includes bibliographic entries, which allows for a detailed description of various categories of source references in term entries.

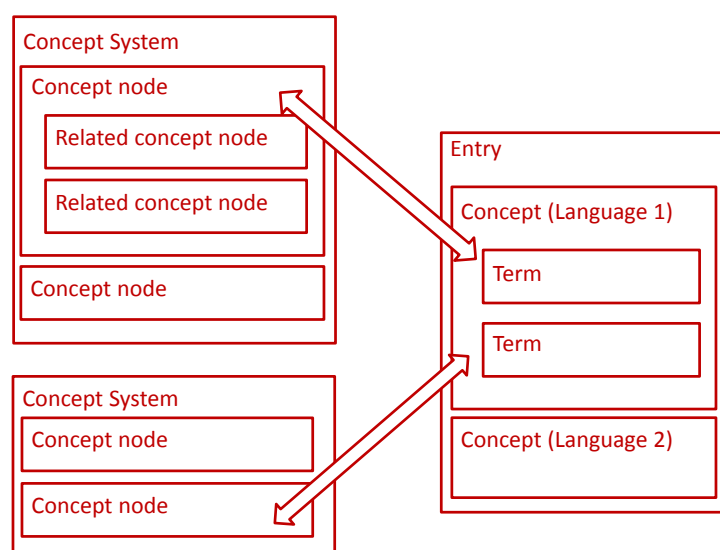
The structure of the terminological entries corresponds to the structure found in TBX (ISO 30042:2008) which is based on the meta-model in TMF (ISO 16642:2003): the Entry Section corresponds to the Entry in the DANTERM Model, the Language Section corresponds to Concept related information, and the Term Section corresponds to Term related information.

3 And then: The relational data base

In the 1990's, the model was implemented in a data base application in the relational data base system Microsoft Access. This application was called DANTERM^{CBS}, and is described in Hull, Madsen & Thomsen (1998) and Madsen and Thomsen (1998).

The database structure included separate tables for all groups of information which could be repeated, thus also separate tables for concept system, position of a concept in one or more concept systems and concept relations. In the original model, these information categories were included in each entry. The introduction of separate tables enabled the generation of systematic lists of entire concept systems to be presented to the user. Figure 1 illustrates the separation of concepts and concept systems. Each node in a concept system is related to a concept, and each concept can be related to a node in one or more concept systems.

Fig. 1. Information about concept systems and related information in the DANTERM^{CBS} data-base structure



4 Later: Extension for terminological ontologies

From 1998-2007 the basic principles of terminological ontologies were developed in the CAOS Project which aimed at semi-automatic development and validation of ontologies (Madsen, Thomsen & Vikner (2004), Madsen & Thomsen (2006)). The principles of terminological ontologies are based on the formalization of concept characteristics according to typed feature theory, c.f. Carpenter (1992). They imply a number of specific constraints which aim at ensuring consistent ontologies and thus a consistent representation of a given domain of knowledge. In terminological ontolo-

gies, characteristics are represented as formal feature specifications, i.e. attribute-value pairs. Subdivision criteria, that have been used for many years in terminology work, were formalized by introducing dimensions and dimension specifications, and these form the basis for the facilities for semi-automatic construction of ontologies and for consistency checking.

When developing the data structure for the CAOS prototype it was necessary to introduce new data categories and data structures. The new data categories were also adopted in ISOcat (ISOcat team, n.d.), the Data Category Registry of ISO TC 37, see table 1.

Table 1. Data categories related to terminological ontologies in ISOcat

Data category	Definition
feature specification	a formal specification of a characteristic of a concept by means of an attribute-value pair
attribute in a feature specification	a part of a feature specification which specifies the feature name
value in a feature specification	a part of a feature specification which specifies the content of an attribute
dimension	an attribute whose possible values allow a distinction between some of the subconcepts of the concept in question
dimension specification	the association of a dimension with its possible values

The extended data model which was needed for storing these data categories was implemented in the relational database system Oracle, and extra tables (objects) were created for the new data categories, c.f. Madsen, Thomsen and Vikner (2002).

5 Meanwhile: Adding graphics for concept systems

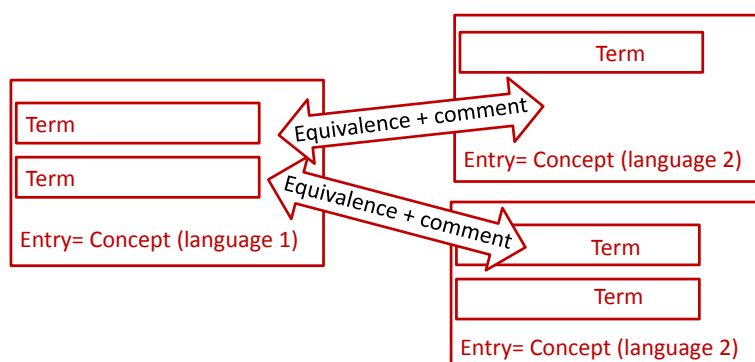
In 2002, DANTERMcentret initiated the development of a web-based terminology and knowledge management system, i-Term[®]. The data fields and the structure of an i-Term entry are based on the original DANTERM Model. In 2004 a graphical ontology module, i-Model, was added to the system. This module allows users to create and present terminological ontologies directly related to the concepts in i-Term, but without the constraints specified in the CAOS project. For example characteristics were implemented as free text without constraints on the form, but the format *attribute:value* was recommended. Subdividing dimensions were presented as subdivision criteria and were only shown in the graphical version of the concept system. The i-Term system is currently used by companies and public authorities in Denmark and other countries.

6 Currently: Revisiting equivalence relations

In 2010, a research project was initiated with the aim of developing a Danish Terminology and Knowledge Bank, and from 2011-2014 the foundations for the DanTermBank were developed with support from the VELUX Foundation. The aim of the project was to develop methods and prototypes for automatic knowledge extraction, automatic construction and updating of terminological ontologies as well as methods for target group oriented knowledge dissemination. For more information and access to the trial term bank, see Thomsen et al. (2016).

One result of this project was the decision to change the basic structure of entries to allow for equivalence relations between one concept in one language and two or more concepts in another language (one-to-many equivalence). As described in Hull, Madsen & Thomsen (1998), Madsen and Thomsen (1998), Thomsen et al. (in print), and Thomsen (2016), this is necessary in many cases of intercultural terminology work. In current termbases, one-to-many equivalence can only be handled by duplicating the entry in the language with one concept, thus creating redundancy in the termbase. The structure needed to handle one-to-many equivalence in a satisfactory way is illustrated in Figure 2.

Fig. 2. The structure needed to account for one-to-many equivalence



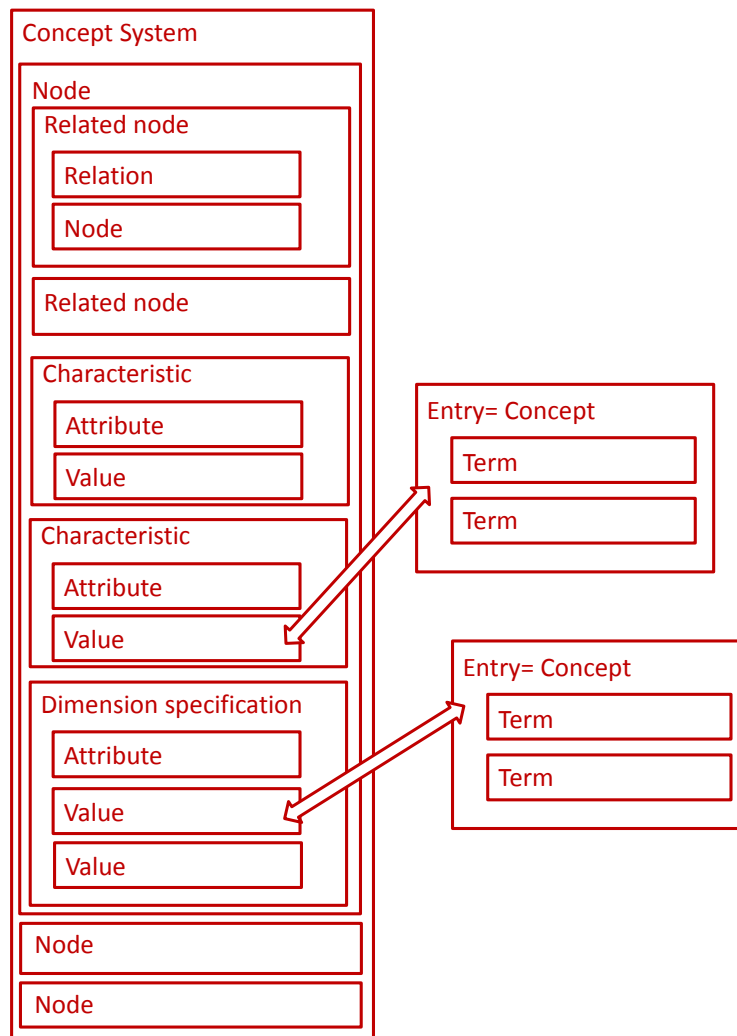
The relational database structure, which formed the basis for DANTERM^{CBS}, already allowed the linking of one concept to two or more concepts without creating redundancy, but the user interface did not exploit this possibility. Current work on the next version of i-Term includes such a change, which will also allow for multi-lingual terminology work in accordance with terminological principles: First terminologists develop concept systems for each language involved, and then, in a second stage, equivalence relations between concepts in two or more languages are established. In the user interface, end users may still be presented with 'entries' containing two or more languages.

We strongly recommend that this new structure is also integrated in future revisions of TBX (ISO 30042:2008) and new versions of ISO standards on terminology databases.

7 The future: Revisiting characteristics and associative relations

Another result of the DanTermBank project was a further extension of the DANTERM Model to encompass all the data categories necessary for handling the constraints on terminological ontologies developed in the CAOS project. This extension is not implemented, but it is necessary in a term bank which includes the automatic consistency checks on terminological ontologies planned for the DanTermBank. The extension involves the information connected to concept systems, where the information on each node or position in the system is specified in more detail as illustrated in Figure 3, where Node, Related node, Characteristic, Dimension specification may be repeated.

Fig. 3. The revised DANTERM model for concept systems



In the new model, characteristics are formalized, they must consist of an attribute and a value, and several characteristics on one node are kept separate. Furthermore, dimension specifications can be added, and these will be related to the characteristics of subordinate concepts.

A more radical revision concerns the treatment of associative relations. In Lassen, Madsen and Thomsen (In print) we argue that the same knowledge can be represented as either a related concept or a characteristic. For example, the knowledge that an α -cell secretes glucagon can be represented either as a relation, *secretes*, between the two concepts *α -cell* and *glucagon*, or as a characteristic, *SECRETES: glucagon*, on the concept *α -cell*. As a consequence, only the type relation and the part-whole relation are to be registered as relations in the future term bank. All associative relations will be registered as characteristics, where the attribute corresponds to the relation, and the value is the related concept, i.e. the value will be another concept in the term-base as illustrated in Figure 3. In a graphical representation of a concept system where both *α -cell* and *glucagon* are included as concepts, the attribute can be shown as a relation, while attribute and value can be shown as a characteristic if only *α -cell* is in the concept system.

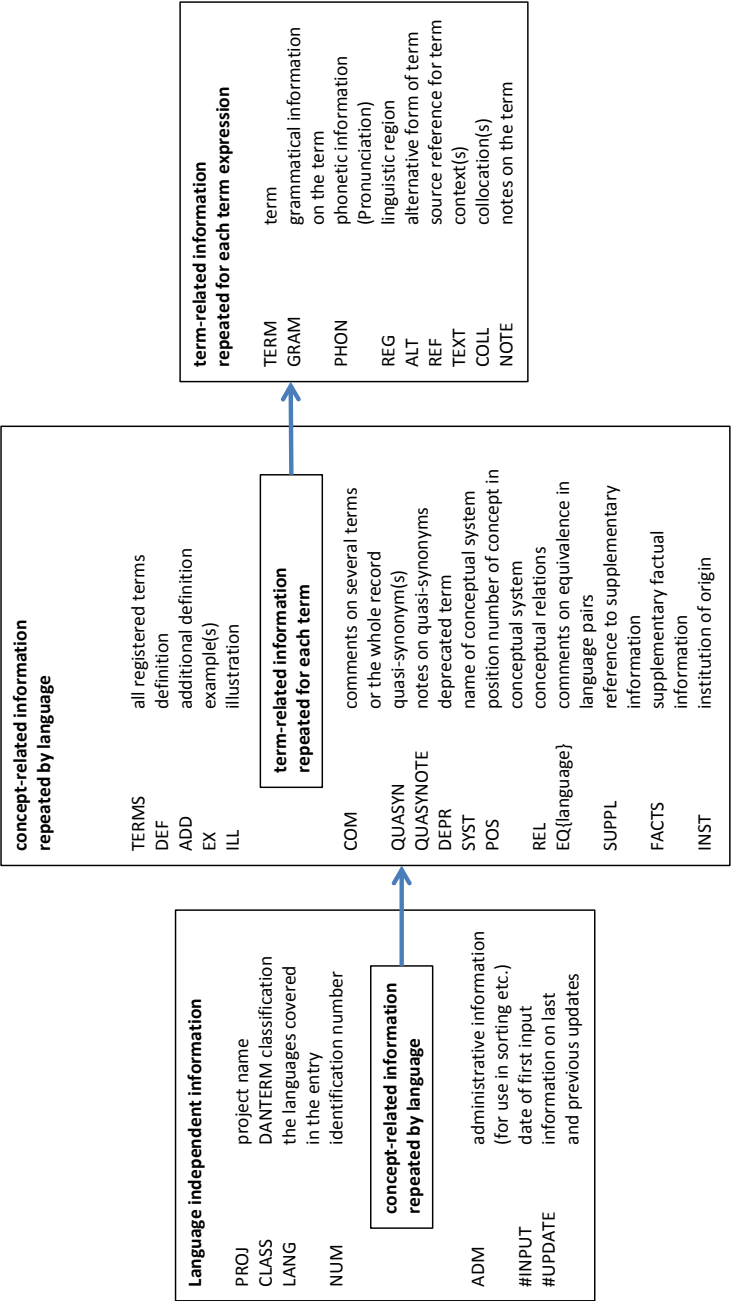
8 Conclusion

The early DANTERM Model developed in the 1970's and 1980's comprised most of the data categories needed for terminology work even today, and also had the basic concept oriented structure that is currently standardized and implemented in most termbase systems. In the 1990's, technological advances, such as the relational data base, enabled the separation of concepts from concept systems in the data base, thereby allowing concept systems to be represented, first in list format, and later on also in graphical form.

Theoretical work on concept systems in the beginning of the new millennium resulted in the need to introduce new data categories for characteristics and subdivision criteria. The most recently proposed revision concerning characteristics and associative relations will enable the use of sophisticated tools for validating terminological ontologies with respect to the inheritance of characteristics.

In the late 1990's, practical multilingual terminology work led us to propose a new structure for handling equivalence, a proposal that we are now putting forward again. The proposed structure with equivalence relations between pairs of concepts, instead of combining equivalent concepts in the same entry, reflects more precisely what equivalence is. It also makes it possible to register equivalence between one concept in one language and two or more concepts in another without having to compromise data base integrity by introducing doublettes. Moreover, it also enables terminologists to work in the prescribed manner, i.e. to register concepts and develop concepts systems for one language at a time and then, in a second step, find equivalence relations. We therefore urge strongly that this change in structure is also introduced in standards for terminology databases and exchange.

Appendix 1: The original DANTERM Model



References

- DS 2394-1. 1998. Collections of Lexical data - Description of data categories and data structure - Part 1: Taxonomy for the classification of information types, (STANLEX), Danish Standards : 74.
- Engel, Gert and Bodil Nistrup Madsen. 1985. DANTERM. In: Heribert Picht, Jennifer Draskau (eds.): *TermNet News 12*. Journal of the International Network for Terminology (TermNet). Special Issue on the Nordic Countries, Infoterm, Wien : 8-10.
- Hull, Anthony, Bodil Nistrup Madsen og Hanne Erdman Thomsen. 1999. DANTERMCBS for everyone. In: *Terminology in Advanced Microcomputer Applications: Proceedings of the 4th TermNet Symposium, TAMA '98*, TermNet, Vienna : 67-84.
- ISO 12620:1999. *Computer assisted terminology management — Data Categories*. International Organization for Standardization.
- ISO 16642:2003. *Computer applications in terminology - Terminological markup framework (TMF)*.
- ISO 30042:2008. *Systems to manage terminology, knowledge, and content – Term-Base eXchange (TBX)*. International Organization for Standardization.
- ISocat team. ISocat - a data category registry. [cited 01.31 2016]. Available from <http://www.isocat.org/>.
- Lassen, Tine, Bodil Nistrup Madsen, and Hanne Erdman Thomsen. In print. Automatisk opbygning og validering af terminologiske ontologier. Paper presented at NORDTERM 2013, Stockholm.
- Madsen, Bodil Nistrup. 1999. The DANTERM Concept. In: *Terminology in Advanced Microcomputer Applications: Proceedings of the 4th TermNet Symposium, TAMA '98*, TermNet, Vienna : 55-65.
- Madsen, Bodil Nistrup, and Hanne Erdman Thomsen. 1998. The DANTERM^{CBS} database. *Kirchmeier-Andersen, Sabine & Hanne Erdman Thomsen (Eds.): LAMBDA Nr. 25 - Datalingvistisk Forenings Årsmøde 1998 på Handelshøjskolen i København 25* : 143-79.
- Madsen, Bodil Nistrup, Hanne Erdman Thomsen & Carl Vikner. 2002. Data Modeling and Conceptual Modelling in the Domain of Terminology. In: Melby, Alan (ed.): *Proceedings of TKE '02 - Terminology and Knowledge Engineering*, INRIA, Nancy: 83-88.
- The DANLEX project group, Danish Working Group on Computational Lexicography (Ebba Hjorth, Jane Rosenkilde Jacobsen, Bodil Nistrup Madsen, Ole Norling-Christensen, Hanne Ruus). 1987. *Descriptive Tools for Electronic Processing of Dictionary Data, Studies in Computational Lexicography*. Lexicographica Series Maior 20, Tübingen, Niemeyer : 285 p.
- The DANTERM project group (Lene Frandsen, Inge Gorm Hansen, Bodil Nistrup Madsen, Jacques Qvistgaard, Gert Engel). 1979. DANTERM - The Danish Terminological Data Bank. *CEBAL no. 5, Special Issue on Terminology*, Handelshøjskolen i København, Erhvervsøkonomisk Forlag : 132-157.

- Thomsen, Hanne Erdman. 2016. Visualizing culture-specific conceptualizations - a tool and a method for concept clarification and intercultural terminology management. Paper presented at Terminologie & Kultur. DTT-Tagungsakte 2016, Mannheim, Deutscher Terminologie-Tag e.V. : 151-160.
- Thomsen, Hanne Erdman, Bodil Nistrup Madsen, and Tine Lassen. In print. Multilingual terminology work in theory – and in practice. Paper presented at Multilingualism in Specialized Communication: Challenges and Opportunities in the Digital Age. Proceedings of the 20th European Symposium on Languages for Special Purposes. Wien.
- Thomsen, Hanne E., Odgaard, Anna E., Madsen, Bodil N., Hoffmann, Pia L. and Lassen, Tine. 2016. DanTermBank - establishing a Danish terminology and knowledge base. [cited 03/20 2016]. Available from www.dantermbank.dk.

Quality Control in Terminology Management

Cristina Valentini

World Intellectual Property Organization, Geneva, Switzerland

cristina.valentini@wipo.int

Abstract. The aim of this paper is to discuss terminology management and the terminology quality control (QC) procedure currently adopted in the PCT Translation Service of the World Intellectual Property Organization (WIPO). First, we will illustrate the structure of the PCT Termbase and the terminology workflow. Second, we will discuss the general principles for conducting terminology QC and the methodology currently in place. Third, we will outline the methodology of semi-automated QC applied to the validated dataset of the PCT Termbase prior to publication in WIPO Pearl, the freely accessible terminology portal of WIPO.

Keywords: quality control, terminology workflow, terminology management, WIPO Pearl.

1 Introduction

As machine translation is becoming ubiquitous on the Web and many Internet browsers offer on-the-fly word disambiguation services that allow users to grasp the meaning of general language words and scientific and technical terms in many different languages, the real added value of a terminology database lies ever more in the reliability and accuracy of the contents. Trustworthy language resources are essential to support the emerging knowledge and content industries as users increasingly do not want to be overburdened with non-evaluated information, but to receive the most pertinent and reliable information for their purposes. For a terminological resource to be as useful possible, it is therefore crucial that mechanisms be put in place to allow monitoring of, e.g., compliance with rules of coherence, use of authoritative sources, linguistic correctness and control of concept redundancy (ISO 23185:2009).

Implementing a quality control (QC) procedure in terminology management is therefore fundamental to ensuring that the information shared with the users is accurate and of high quality. While terminology is a key element in quality assurance and quality control of document production, and particularly translation, there is still scarcely any discussion of practical implementation of quality control procedures of terminological products in the literature (Galinski and Budin 1993, Pozzi 1996, Wright 2001, Kockaert and Steurs 2015).

The aim of this paper is to present the terminology QC procedure currently implemented in the Translation Service of the Patent Cooperation Treaty (PCT) of the World Intellectual Property Organization (WIPO). First, we will illustrate the struc-

ture of the PCT Termbase and the terminology workflow. Second, we will discuss the general principles for conducting terminology QC and the methodology currently in place for contributing and validating terms in the PCT Termbase. Third, we will outline the methodology of semi-automated QC applied to the manually validated dataset of the PCT Termbase prior to publication in WIPO Pearl, the freely accessible terminology portal of WIPO.

2 The PCT Termbase¹

The World Intellectual Property Organization (WIPO) is a United Nations agency and a forum for intellectual property (IP) services, policy, information and cooperation. In particular, it provides access to the world's IP information via its free global databases, such as PATENTSCOPE, the Global Brand Database, the Global Design Database, WIPO Lex and, since 2014, WIPO Pearl.²

The Patent Cooperation Treaty (PCT) is one of the 26 international treaties administered by WIPO. The PCT system makes it possible to seek patent protection for an invention simultaneously in a large number of countries by filing a single international patent application instead of filing several separate national and regional patent applications at the outset. By 2014, the PCT comprised 148 contracting states (WIPO 2015). The PCT Translation Service is responsible for translating into English and French the titles, abstracts and the text in the drawings of international patent applications ahead of their publication, and for translating into English the international search reports, written opinions and preliminary reports on patentability relating to these applications. Approximately 125 million words of translation were carried out in 2015.

The PCT Termbase is currently developed within the PCT Translation Service and includes scientific and technical terms extracted from abstracts and titles of international patent applications filed through the PCT system in the ten PCT publication languages, namely Arabic, Chinese, English, French, German, Korean, Japanese, Portuguese, Russian and Spanish. The termbase also contains IP terms related to patents and to the Patent Cooperation Treaty. The database structure complies with relevant ISO terminology standards (ISO 1087-1:2000, ISO 704:2009, ISO 12620:2009) and includes data categories for recording conceptual and linguistic information.

The building block of the PCT Termbase structure is the record that uniquely identifies a concept. A record has three levels: the Entry Level, the Language Level and the Term Level. The Entry Level gives information pertaining to the concept. The Language Level helps to situate the concept within a given language and includes data categories for indicating potential differences in concept coverage across the various languages. The Term Level gives information on the nature and status of preferred terms and synonyms entered in the record. In particular:

¹ The views expressed in this article are those of the author and do not necessarily reflect the views of the World Intellectual Property Organization.

² <http://www.wipo.int/reference/en/wipopearl>.

- At the Entry Level, the concept is assigned to a unique subject field and related subfield selected from among the 29 subject fields and 311 subfields available in a specially devised classification that represents the backbone of the dataset.³
- Hierarchical and non-hierarchical concept relations are also established at the Entry Level and are valid for all the languages comprised in the record.
- The record includes designations for the concept in question in at least 2 and at most 10 languages, with defining contexts for the concept entered in each language.
- For each concept in each language, a preferred term is identified, establishing a hierarchy among the possible designations.
- A reliability code is attributed to each term block that fully complies with internal terminology guidelines.⁴
- Additional information such as recommendation on term usage can be entered by completing one of the additional data fields specifically devised to record such information at the Term Level, namely Usage Label and Term Description.⁵

As of September 2014, the contents of the PCT Termbase have been published online, in WIPO Pearl. The development of the web interface was intended to provide different users alternative ways of accessing the information contained in the PCT Termbase according to their specific needs and in relation to other services already available on the WIPO website. For example, WIPO Pearl is linked to PATENTSCOPE, WIPO's patent corpus, and two other specific tools embedded in it, namely PATENTSCOPE CLIR, the Cross-lingual Information Retrieval tool, and WIPO Translate, the internally developed patent-trained machine translation engine, formerly known as TAPTA (Translation Assistant for Patent Abstracts and Titles) (Pouliquen and Mazenc 2011).

In general, the integration of the terminology database with the documentary database allows users to find additional contexts of use for the validated terms published in WIPO Pearl. Moreover, it offers patent stakeholders the option to search for patent applications published by the PCT and other regional and national offices. By contrast, machine translation is used to offer suggestions for a queried term existing in the PCT Termbase but for which one of the selected target languages is missing. If the queried term does not exist at all in the PCT Termbase, the user can also launch PATENTSCOPE CLIR, the cross-lingual information retrieval tool that will machine-

³ See WIPO Pearl Concept Map Search Interface for an overview of all subject fields and related subfields: <http://www.wipo.int/wipopearl/search/conceptMapSearch.html>.

⁴ A term block comprises the term itself, the context, source, usage label, and any other term-base field that may have been filled in.

⁵ The Usage Label and Term Description are two specially devised data fields in the PCT Termbase that correspond to the description of the data category "Term Type" in ISO 12620:2009. The Usage Label includes values such as *allowed*, *avoid*, *obsolete*, *proposed term*, *recommended* and *standardized*. The Term Description field includes values such as *abbreviated form*, *chemical name*, *common name*, *formula*, *full form*, *generic name*, *geographical variant* (and its nested field, Variant Code), *scientific name*, *spelling variant* and *unit*. See Rouquet et al. *in print* for a full overview of the PCT Termbase structure and a discussion of the different data categories used.

translate the term in question into the target language(s) selected and provide evidence of use of such translation proposals in PATENTSCOPE.

Finally, WIPO Pearl innovatively includes a concept map search interface that allows the users to browse and search the dataset by concept and analyze the relations existing between concepts in a specific subfield.

3 The PCT Terminology Workflow

Concepts and terms are contributed daily by PCT staff terminologists, translators, and short-term terminology trainees. Contributions may involve creation of new records for concepts not existing in the PCT Termbase or completion of existing records by adding missing languages. Term extraction is used to populate the PCT Termbase with new concepts, as well as to identify existing concepts for which designations in certain languages do not yet appear in the PCT Termbase and should be entered as a priority. Each term contributed is then assigned to a validator - typically another fellow terminologist and/or translator, and ideally a native speaker of the language of the term in question.

Terminology validation involves confirming that a record or term that has been entered accurately reflects the expression of a single concept and that the term in a given language is indeed the most accurate designation for that concept. In addition, validation also ensures that the content of each field is formally consistent with the principles established in internal terminology guidelines. Further, when terms are switched from “candidate” to “validated”, a term reliability score is assigned according to a scale of 1 to 4. No term block is published in WIPO Pearl until it has been awarded the status “validated”, the only exception being WIPO MT results. The latter are, however, clearly identified in WIPO Pearl results’ list as translation proposals with no term reliability score assigned.

4 Principles of Terminology Quality Control

Maintaining the necessary level of quality is paramount to delivering a reliable terminology product to (i) internal and external PCT translators, so as to improve the overall quality/consistency of the translations delivered, and (ii) the public at large, so that WIPO is perceived as a reliable provider of multilingual scientific, technical and IP terminological information.

A terminology database is assumed to be in compliance with the quality standards when it is “fit for purpose”, i.e. is suitable for the end use for which it is intended (EN 15038:2006). Generally speaking, a terminology record is considered to meet the required quality standards when the information it contains is complete and reliable in such a way that end-users do not need to verify it further using additional resources, i.e. the record is self-sufficient and users can find in it all the information they need to (i) understand the underlying concept, (ii) trust the equivalence given for a term in a given language.

The key criteria identified when assessing the quality of terminology work are namely relevance, completeness, correctness and informativeness.

4.1 Relevance

Relevance is determined by assessing whether a concept or a term is suitable for inclusion in the PCT Termbase. According to the internal terminology guidelines, concepts are deemed relevant if they are key concepts and/or belong to the state of the art technology in one of the subject fields of interest for patents. For instance, in English, relevant key terms in the field of optoelectronics are “laser”, “wavelength” and “semiconductor”, while an example of state of the art technology terms are “FOLED”, “WOLED”, “PMOLED” and “AMOLED” that designate recently developed types of organic light emitting diodes (OLED).

On the other hand, terms are deemed not suitable if, for example, they are general descriptive terms, lone adjectives, appellations or trademarks. In particular, general descriptive terms represent a specific challenge in patents as patent drafters tend to use general language words in nominal phrases to refer collectively to classes of objects for which a standard designation does not exist, in order to broaden the scope of protection for their invention. In English, this is seen in nominal phrases that include general-purpose words such as, for example, ‘means’, ‘device’, ‘substance’, ‘apparatus’, ‘equipment’, ‘material’, ‘medium’ or ‘element’.

4.2 Completeness

A record is deemed complete when all mandatory fields are filled in at the Entry, Language, and Term Level. Mandatory fields in the PCT Termbase structure are Subject field, Subfield, Original Entry Language (the language that prompted the creation of the record), Term, Term Status (*candidate*, *validated*, *unresolved*), Term Reliability (1-4), Usage Label, Term Type (*head term*, *synonym*), Context and Source.

4.3 Correctness

Correctness is determined by assessing primarily whether the content is substantively correct, including correctness of term equivalence, subject field/subfield, synonyms, contexts, term notes, reliability scores assigned to terms, etc. A second criterion is formal correctness, i.e. spelling/grammar, punctuation, appropriate form of terms, sources and references.

4.4 Informativeness

Content is also assessed with regard to informativeness, namely the extent to which the end-user can understand the concept and usage of a term in a specific subject field. This involves checking whether, for example, the context contains a definition, and the sources are credible, i.e. written by native speakers, derived from patents or

scientific and technical articles or textbooks, if any additional useful information on the concept has been provided in the optional fields.

The Context field, in particular, is used instead of the Definition field to provide information about the concept and evidence of use of the term in question in the specific subject field and subfield selected. For instance, for “wireless network” an appropriate context would be in the field of computer networks. Moreover, the context should ideally contain a definition or some items of knowledge that could be useful to understand the concept following Meyer’s definition of “knowledge rich-context” (Meyer 2001). Examples (1) and (2) below contain therefore good contexts, whilst the content in example (3) would not be entirely appropriate as the context is merely associative and deals with urban infrastructure building:

(1) “The term wireless network generally refers to a telecommunications network whose interconnections between nodes are implemented without the use of wires.” WO/2008/121974

(2) “As wireless technology has advanced, a variety of wireless networks have been installed, such as cellular, wireless LAN (local area network) or WLAN, and other wireless networks. Some wireless networks are based upon the Institute of Electrical and Electronics Engineers (IEEE) 802.11 family of Wireless LAN (WLAN) industry specifications, or other IEEE specifications, for example. Other wireless networks are based on cellular technologies, such as Global System for Mobile Communications (GSM), for example. Some networks are being developed based on other standards or technologies, such as WiMedia ultra-wideband (UWB) common radio platform to augment the convergence platform with TCP/IP services.” WO/2008/035161

(3) “Networked ISL lamp communicates with the central system ISH (2D) via a wireless network (2B) and Internet (2C) delivering all collected data related to atmospheric conditions at the microlocation as well as ISL status information.” WO/2013/019135

5 Terminology QC Methodology

Terminology Quality Control is currently conducted in two different ways:

- manual QC of contributions
- semi-automated QC of all validated termbase content.

The manual QC procedure is currently implemented for some of the candidate terms when performing validation of contributions. The main aim is to check substantively whether the contents of the candidate term blocks comply with the aforementioned criteria of relevance, correctness and informativeness. Furthermore, it also helps monitor the progress of certain contributors by evaluating their work.

In particular, the following categories of errors are considered when assessing the quality of a candidate term block:

1. Term
2. Context
3. Source
4. Mandatory fields
5. Data integrity
6. Proofreading

Data integrity is a particular category in database management and can be defined as the assurance that the data is correct and consistent in relation to a pre-established set of acceptance criteria. Such criteria are defined in the internal terminology guidelines and involve checking whether certain types and forms of data elements are allowed for a certain data category (Schmitz 2001).

The rating system for terminology QC is based on the distinction between major and minor errors and the general principle is acceptability for publication in WIPO Pearl. Major errors are significant inaccuracies or significant omissions in key term-base fields. Examples of major errors are:

- Term: the term is the wrong designation in that language for the concept, or the term is not suitable for inclusion in the PCT Termbase. For instance, the term denotes an individual concept, i.e. is an appellation. Examples are: names of operating systems and programming languages such as Windows and JAVA.
- Context: the context does not contain the term or the term appears as part of a larger multi-word unit (i.e. the context does not contain the term in isolation). To exemplify, contexts for “wireless network” that contain “wireless network *service*”, “wireless network *configuration*” or “*ad hoc* wireless network” would not be suitable, since “wireless network” appears in compounds that may actually designate different concepts.
- Source: the source is not eligible. Examples of sources that should be avoided are: dictionaries, WIKIPEDIA and online forums. Moreover, a source is not eligible if it is a translation.
- Mandatory fields: mandatory fields as described in section 4.2 are not completed.
- Data integrity: contents are entered in the wrong fields, including when information for which a specific option can be selected from a picklist is provided in free-text fields (Term Note, Context, Source). An example is adding information on geographical usage in the Term Note field instead of selecting “geographical variant” from the picklist of the data category Term Description and the appropriate ISO language code from the nested Variant Code field.
- Proofreading: the term is misspelled or the source is given incorrectly so that it cannot be retrieved, e.g. wrong patent number, wrong title of book.

Minor errors are less serious in impact than major errors, but diminish the overall reliability of the termbase and the likelihood that users will regard it as a reference. Minor errors include typically data integrity and proofreading errors such as incorrect selection of picklist values and free-text fields containing spelling or grammatical errors.

6 Semi-automated QC of Terminology Validation

Semi-automatic quality control (QC) is performed on all manually validated term blocks with regard to the criteria of completeness, formal correctness and respect of terminology work procedures such as avoidance of self-validation. Errors are typically detected on the basis of the incorrect association of values assigned to certain fields within a record or erroneous multiple selections of values within a term block. The procedure is termed “semi-automatic” because, in some cases, once the record and error in question have been identified by running an automated script on an XML extract of termbase contents, a further manual check is needed.⁶

Semi-automatic QC does not involve an assessment of the validated contents with regard to relevance and other aspects of substantive correctness, e.g. it does not assess the suitability of the term block for inclusion in the database or the accuracy of term equivalence, subject field or subfield selected, or quality of contexts. These aspects are, however, checked on an ad hoc basis in the course of procedures such as merging of records, or indeed when term blocks containing errors identified by semi-automatic QC are examined and corrected.

As mentioned above, the distinction between major and minor errors is based on the extent to which the error can affect the upload and display of data in WIPO Pearl. In the Concept Map Search, for example, information entered in the Subject field and Subfield data categories is used to browse the database. Thus, an error in the association of subject field and subfield can jeopardize the use of the search tool, and is therefore deemed a major error. Some of the other aspects monitored are exemplified in table 1 below.

Table 1. Examples of errors monitored in semi-automated QC.

<i>Data category</i>	<i>Description</i>
Related Concepts	Records in which the Related Concept, Related Concept Broader or Related Concept Narrower fields has been completed but reverse links to other record numbers in the PCT Termbase are incorrectly entered
Term Reliability	Records in which the Term Status value is <i>validated</i> and the Term Reliability score is 2 or 4, or has not been completed
Usage Label	Records in which, within a term block, the Usage Label is <i>proposed term</i> and the Context field has been completed, or records in which, within a term block, the Usage Label value is <i>proposed term</i> and the Term Reliability score is 3

⁶ Scripts are run each week on an export of the termbase. The weekly QC statistics tool has been designed within the PCT Translation Service and is maintained internally.

Term Type	Records in which, within a language block, the Term Type value <i>head term</i> has not been selected or has been selected more than once
Term Description	Records in which, within a term block, the Term Type value is <i>head term</i> and the Term Description is <i>spelling variant</i>
Context	Records in which, within a language block, the Term Type value is <i>synonym</i> and the Context field has not been completed

The incidence of errors detected after validation is low. The major errors identified typically relate to incorrect completion of fields for recording concept relations, Term Reliability scores, association of Subject fields and Subfields, Usage Label and Term Description values, such as multiple selection or non-selection of a head term value. The majority of these errors could be avoided if mainstream terminology management systems offered data validation options such as a system of automatic checks for matching certain data field values with others (Schmitz 2001).

Semi-automated QC errors are tracked weekly and feedback is given to the validators. Although the QC procedure implemented by means of the semi-automated checks can be generally considered as a sort of formal data validation procedure, it often also allows substantive errors to be identified that reveal a gap in the knowledge or understanding of terminology principles. Thus, it represents an important tool for bringing to the fore specific training needs of staff performing terminology work that can subsequently be addressed in specific training sessions.

7 Conclusion

Quality control is a crucial aspect of terminology management especially for terminology databases that are shared on the Web and/or may be leveraged in other applications, such as semantic search and information retrieval systems. In this paper, we have discussed the QC procedure implemented for the PCT Termbase, whose contents are published regularly in the online terminology portal, WIPO Pearl. A manual QC of terminology contributions performed during the validation phase and a semi-automated QC of all validated data are implemented in order to ensure that key quality criteria of relevance, completeness, correctness and informativeness are achieved.

Advantages of implementing a QC procedure such as the one described in this paper are many, amongst which we can highlight: increased reliability of terminology resources, enhancement of terminology workflow management procedures, and identification of terminology training needs in the group of collaborators. In this scenario, relevance and completeness are the most difficult aspects to monitor from a substantive viewpoint in a systematic way. In order for terminology QC to be effective, it needs be (i) carried out regularly on a specified set of data that is delivered within a certain timespan and be representative of the content of the termbase, (ii) performed ideally by native speakers of the language in question who are language experts/terminologists. However, this sometimes proves difficult because of the many different language combinations covered by the PCT Termbase and because of the often limited time that can be dedicated to the different terminology tasks.

8 References

EN 15038:2006 Quality Management for Translation Service Providers

Galinski, Christian, and Gerhard Budin. 1993. “Comprehensive Quality Control in Standards Text Production and Retrieval” in *Standardizing Terminology for Better Communication: Practice, Applied Theory, and Results*, edited by Strehlow, Richard, A., and Sue E. Wright, 65-74. Philadelphia: American Society for Testing and Materials.

ISO 1087-1:2000 Terminology Work – Vocabulary – Part 1

ISO 704:2009 Terminology Work – Principles and Methods

ISO 12620:2009 Terminology and other language and content resources

ISO 23185:2009 Assessment and benchmarking of terminological resources – General concepts, principles and requirements

Kockaert, Hendrik, and Frieda Steurs, eds. 2015. *Handbook of Terminology, Volume 1*. Amsterdam: John Benjamins.

Meyer, Ingrid. 2001. “Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework” in *Recent Advances in Computational Terminology*, edited by Bourigault, Didier, Jacquemin, Christian, and Marie-Claude L’Homme, 279-302. Amsterdam: John Benjamins.

Pouliquen, Bruno and Christophe Mazenc. 2011. “Coppa, CLIR and TAPTA: three tools to assist in overcoming the patent barrier at WIPO”. Paper presented at the MT Summit XIII: the Thirteenth Machine Translation Summit, Asia-Pacific Association for Machine Translation (AAMT), Xiamen, China 24-30.

Pozzi, Maria. 1996. “Quality assure of terminology available on the international computer networks” in *Terminology, LSP, and Translation*, edited by Somers, Harold, 67-82. Amsterdam: John Benjamins.

Rouquet, Philippe, Valentini Cristina, and Geoffrey Westgate. In print. “The PCT Termbase of the World Intellectual Property Organization: Designing a database for multilingual patent terminology”. *Terminology*.

Schmitz, Klaus-Dirk. 2001. “Criteria for Evaluating Terminology Database Management Programs” in *Handbook of Terminology Management, volume 2*, edited by Sue Ellen Wright and Gerhard Budin, 539-552. Amsterdam: John Benjamins.

World Intellectual Property Organization. 2015. *Patent Cooperation Treaty Yearly Review*. Geneva: World Intellectual Property Organization.

Wright, Sue E. 2001. “Terminology and Total Quality Management” in *Handbook of Terminology Management, volume 2*, edited by Sue E. Wright and Gerhard Budin, 488-503. Amsterdam: John Benjamins.

Web Interfaces of Terminological Databases that are Available on the Internet from a Usability Perspective

Barbara Heinisch-Obermoser

University of Vienna, Centre for Translation Studies, Vienna, Austria

`barbara.heinisch-obermoser@univie.ac.at`

Abstract. The usability of termbases has received considerable attention in the research literature. Some organizations make their termbases publicly available on the Internet to a broader user group. However, little is known about the usability of these termbases. This paper analyzes the web interfaces of eight termbases from a usability perspective. The study seeks to explain the usability qualities of these termbases by relating a set of tasks and the resulting screens with latest findings in termbase usability research. The results show that the majority of the analyzed termbases focus on the users' previous experience with similar systems and reduce the complexity of their web interfaces.

Keywords: Usability. Terminological database. Termbase. Terminology management system. Web interface.

1 Introduction

There is a growing body of literature that recognizes the importance of usability of human-computer interfaces. Usability also becomes an area of interest within translation studies including terminological database design (e.g. Marcos et al. 2006; Sevriens 2010) and computer-assisted translation tools (CAT tools) (e.g. Höge 2002; Tuominen 2012). To make a terminological database usable to as broad a target group as possible usability aspects should already be considered during the design phase of a termbase's web interface.

Usability can be a result, a process, a set of techniques or the philosophy of designing (Quesenbery 2001). Although there are slightly different definitions of usability in the literature, it usually consists of seven elements: the product, its users, the users' goals, effectiveness, efficiency, satisfaction and the context of use (Go 2009, 195). Usability describes the ease of a system's use. If users do not experience frustration while using a product or service, it is regarded as usable (Rubin and Chisnell 2008, 4). The five quality components that are usually related to usability are learnability, efficiency, memorability, error tolerance and easy error recovery, and satisfaction (Nielsen 2010, 26). Interactive systems should follow the ergonomic principles of suitability for the task, suitability for learning, suitability for individualization, conformity

with user expectations, self-descriptiveness, controllability and error tolerance (ISO 9241-110:2006).

Previous research has found that usability heuristics applied to terminological databases (termbases) include a navigation that facilitates information retrieval, specified functions of the database (such as languages, domains or user groups), user control, written material intelligible to the target group, online help and user guidance, system feedback, accessibility, consistency (of graphic design), error prevention and architectural clarity (Marcos et al. 2006).

2 Usability Aspects of Web Interfaces of Termbases Available on the Internet

2.1 General Aspects of Termbase Usability

Termbase Design. The first steps in termbase design consist of the definition of activities, roles, tools, workflows and means of cooperation and communication (Chiocchetti and Ralli 2013). This means that usability principles should also be considered in termbase creation. These principles influence a database's layout, design, workflows, data selection, data compilation and data display as well as the selection of entry models, etc. Termbase interfaces should be user-centered, i.e. the user's point of view is crucial so that users can perform actions with the termbase the way they expect to do it.

User and Needs Analysis. Understanding the users and their needs is of pivotal importance in usability engineering. Information on the user groups helps defining the complexity and content of the user interface (Nielsen 2010). The needs of the user groups influence the selection of (visible) data, the amount of information provided, the structure of the data, the layout of the termbase, the options for the search functionality, etc. The user groups of termbases comprise terminologists, translators, domain experts, company employees, etc., and the general public if the termbase is freely available on the Internet. These user groups include people from different professional and educational backgrounds, of different age, gender and with different levels of previous experience of computers or software applications. In addition, there are always individual differences between users.

Generally, a wide range of user groups should benefit from the same design of a system (Nielsen 2010, 43). Based on a specified terminology workflow different roles such as terminologists, approvers or translators may require different user interfaces because they complete different tasks and need different system functions. The more sophisticated the workflow and the more tasks a role has, the higher the number of functions in a termbase's user interface. However, this makes a user interface more complicated and decreases a system's learnability. If the workflow provides for collaboration with users who are not familiar with terminology management, the user interface has to be as simple as possible.

The web interface and content of a terminological resource that is made available on the Internet to the general public might differ from the interface and content available to other roles. Hence, users might not be able view terms that are not yet approved or validated in a termbase's public version. Moreover, some terminological data or termbase functions might be hidden to reduce the complexity of the interface. This makes a system easy to learn.

Users rely on their previous knowledge of other systems and transfer this knowledge to new systems (Nielsen 2010, 45–46). This means that the users' expectations derived from similar systems such as online dictionaries or search engines extend to termbase use as well. As these systems have become easy to use, users also expect to be able to use termbases without prior training or without consulting help or documentation. Therefore, the web interface of a termbase should reflect the features that users are already familiar with, e.g. information search and information display.

The users' knowledge of a system's domain is also important. Domain experts are familiar with domain terminology. Therefore, domain terminology can be used and the information density on the user interface can be higher (Nielsen 2010, 46).

Website. The look and feel of a termbase's web interface can improve user experience and satisfaction. A termbase embedded in a web environment should follow general web usability guidelines related to information architecture, page layout, graphic design, writing for the web (Brinck, Gergle, and Wood 2002), accessibility and responsive design. The design and content of the interface should be consistent with the organization's corporate identity. In addition, a localized version of a termbase's web interface and technical documentation can also enhance user experience.

Content. The scope and quality of terminological data are major factors that determine the usability and credibility of a termbase. Therefore, terminology managers usually focus on the content of termbases, i.e. the maintenance and structure of a termbase, the selection of data fields, the compliance with terminology standards, interoperability and correct spelling of written material. However, in terminology management the presentation of this content is often of secondary importance. Nevertheless, usability considerations do not only influence the termbase layout but also the selection and preparation of a termbase's content. Regarding content, termbase designers only select those domains, data categories, data elements, etc. that are relevant to their (primary) user group. Concerning termbase layout, termbase designers might want to draw attention to some data (e.g. by using a larger font or color for these data) or divert attention from other data (e.g. by using grey color for deprecated terms). Usability might also require that some items such as definitions are (re-)written according to the needs, tasks and domain knowledge of the users.

The designations of the fields on the web interface should be intelligible to various user groups and avoid misunderstandings (Lemmetti 2001, 85). Therefore, it might be necessary to change the designations of data fields according to the users' needs.

Consistent data and linguistic correctness of the information provided increase the credibility of the termbase and its content. This is also the reason why many term-

bases that are available on the Internet only display approved terms in the result list. However, not only the content of a termbase but also its functionalities as well as its look and feel enhance its credibility and usability.

2.2 Analysis of the Web Interfaces of Publicly Available Termbases on the Internet

A sample of eight termbases that are publicly available on the Internet was analyzed for usability attributes. Criteria for selecting the termbases were as follows: They are freely available on the Internet, provide a web interface for searching terms in more than two languages and enjoy a certain degree of popularity. Therefore, the termbases analyzed were EuroTermBank, FAO TERM PORTAL, IATE, Microsoft Terminology Collection, SAPterm, TERMIUM Plus®, UNTERM and WTOTERM. The analysis was based on a set of tasks to be completed with these termbases including finding an English term and an equivalent term in another language, finding a term's definition, browsing the termbase's content, commenting on an entry's content or requesting a new term, sharing an entry, integrating the termbase into other tools and searching a term that is not in the termbase to elicit an error message. Table 1 provides an overview of selected functions and the following section summarizes the functions offered by these termbases for completing the aforementioned tasks and relates them to latest findings in termbase usability research.

Screen Layout. The presentation of information in a termbase influences the effectiveness of the end-user (Cauna 2012). Therefore, web interfaces of termbases should follow web design guidelines. This includes a consistent interface that provides concise information. The search and results screen of the studied termbases' web interfaces share the same screen layout, i.e. the same information is displayed at the same location on the screen and the format is consistent across all pages. Moreover, except for two of the analyzed termbases users can access the search box on all screens. This consistency is an important usability attribute that increases the user's confidence in using the system. The majority of the termbases keep navigation to a minimum and their search field is reduced to one large central bar in the interface. In most cases, the search field is clearly distinguishable from other sections on the screen. This differentiation between information items is especially important on the results screen because it displays terms in various languages and additional data such as definitions, sources or notes. WTOTERM provides a selection of different layouts for displaying the termbase's content. This leads to more flexibility. Moreover, users who just want to get an overview of a list of terms or a terminological entry are not intimidated by a confusing interface and a lot of information on the screen.

Some start screens of the termbases studied provide information on the content of the termbase, i.e. information on the organization responsible for the content, the domains and languages addressed or the number of terms available. This information guides the users' expectations and enhances the users' efficiency in using the system.

Table 1. Selected functions of the analyzed termbases according to the completion of the tasks described in section 2.2.

Feature / termbase	Euro-TermBank	FAO TERM PORTAL	IATE	Microsoft Terminology Collection	SAPterm	TERMIUM Plus®	UNTERM	WTO-TERM
Consistent screen layout	yes	yes	yes	yes	yes	yes	yes	yes
Search form on both search and results screen	yes	yes	yes	yes	no	yes	yes	yes
Search query (auto-complete)	no	yes	yes	no	no	no	yes	no
Different search modes (excl. filters, incl. wildcards) on search screen	yes	yes	yes	no	yes	yes	yes	yes
Thematic search	no	no	no	no	no	no	no	no
Result list (refining results without new search)	yes	yes	no	no	(no)	yes	yes	(no)
Terminological entry (containing hidden fields)	depending on display options	no	no	no	no	depending on display options	no	depending on layout
Help and documentation (incl. FAQs, tooltips)	yes	yes	yes	no	yes	yes	yes	yes
Error messages (providing solutions)	no	yes	(yes)	no	no	yes	yes	no
Individualization	no	no	yes	no	no	yes	yes	no
Direct integration into CAT tools	no	no	no	no	no	no	no	no
Feedback options (accessible from the screen)	yes	yes	yes	yes	no	yes	yes	no
Social features (share entry)	no	yes	no	no	no	yes	no	no

Search Form, Search Query and Search Modes. The search function is a central feature for termbase users (ISO 26162:2012). Users prefer a central search field, concise information and an overview of all the features available in a termbase (Bank 2012, 359). They want to access the information they are looking for quickly and efficiently. Therefore, the main characteristics of publicly accessible termbases are a web interface of reduced complexity and many ways to achieve the same outcome. The majority of the termbases analyzed have a distinctive and large search field that is placed at a prominent location on the web page and is accessible from all screens. As the search function of the studied termbases is not case or accent sensitive users can search efficiently. Some termbases also offer localized versions of their web interface.

The sequence of the fields visible on the search screen should reflect the natural sequence of the steps that have to be taken when searching a term. The majority of the examined termbases follow a logical sequence as the first field requires the users to enter a search query and the second field to select other fields, e.g. search mode or languages. The search forms are either placed in one horizontal search bar or are arranged vertically.

None of the analyzed termbases allows users to conduct a thematic search or browse an ontology for getting an overview of the terms available in a certain domain. However, two termbases provide alphabetical lists of their terms.

The majority of the analyzed termbases offer different search modes. The basic distinction is between simple and advanced search. The simple search is a more general search that offers only a small number of search options to select from or no search options at all. Basically, the simple search of the analyzed termbases includes a search box for entering the search query, (fields for selecting the source and target language or a combination of target languages) and a search button. Some termbases also offer a small number of additional search options or filters for the simple search. Those termbases that do not offer a language selection option on the simple search screen reduce the complexity of the user interface but decrease the controllability of the system. Termbases that only have a limited number of search options on the search screen allow for the refinement of the results on the results screen. This enables users to recover from possible errors. Reasons for making the simple search mode the default search mode are the high learnability, i.e. novice and casual users can quickly (learn to) use the system because there are only a few features to use.

Some of the termbases offer both a simple and an advanced search on the search screen. The advanced search allows users to search across various termbase fields in order to get more precise results. Thus, it increases the users' control. The majority of the termbases examined offer various forms of exact match, fuzzy match and free-text search (see ISO 26162:2012). To broaden or refine a search some of the analyzed termbases support Boolean operators or wildcard characters. Termbases that offer an advanced search option display a button or hyperlink to access the advanced search from the start search screen and enable users to search in certain domains and/or choose a search mode. In some cases, users can change the search mode not only on the search screen but also on the results screen.

The most common mistakes that users make when searching in a termbase are the choice of the wrong search language(s), including the wrong language direction, spelling errors in the search query and searching for terms that are not in the termbase. The consequence of these mistakes is that the search does not yield results. Language suggestion and auto-complete functions are functions that can prevent these errors from occurring or help termbase users recover more quickly. A language suggestion function supports users in choosing the correct target languages. Here the interface displays similarly written terms in another language and proposes them to the users. Alternatively, auto-complete can suggest terms while users enter the search query in the search box. The list of terms that are suggested by auto-complete can be sorted alphabetically, ranked according to their popularity derived from previous search queries or according to the most recent record. Auto-complete reduces spelling errors and gives users an overview of terms available in the termbase. Furthermore, users can enter their search query more efficiently and accurately and they see if they have chosen the wrong search language (Sevriens 2010, 40-63). However, only three of the examined termbases have an auto-complete function while entering the query. If terms are entered in the wrong source language, many of the studied termbases do not suggest similar terms in another language but prompt users to suggest a new term.

Users should get appropriate system feedback in reasonable time. This means that the response time of the system should be low. If the system does not respond immediately, users start doing other tasks and are not attentive any more. Therefore, the system should provide information on its status. Some of the examined termbases have progress indicators such as spinners while processing the search to provide system feedback to the users.

Result List and Terminological Entry. The result lists of the analyzed termbases display all results that contain the search query (according to the search mode used), i.e. relevant or similar entries or terms. In most cases, the results screens provide either a list of all relevant entries found in the termbase or directly display the best-matching terminological entry or entries. These lists are either displayed in a separate column or at the center of the screen. The majority of the termbases analyzed display only a limited number of terms in the result list. Thus, users do not have to scroll down a long list of results and are more efficient in using the termbases. The systems display either entries in the language combination selected on the start screen or return results from any language combination if a simple search (without selecting language combinations) was conducted. The terms per language are either displayed horizontally one below the other or vertically in columns, e.g. one column per language. As the term is the crucial information item for users, many termbases highlight the term of the search string or the relevant language headings on the results screen, e.g. by using a bold font style or different color to make them clearly distinguishable from terms in other languages or other terminological data such as definitions.

To reduce the number of entries in the result list some termbases allow users to narrow their search. On the one hand, users may refine and narrow their search by using filters or by sorting results according to certain criteria such as language(s) or domain. On the other hand, users may define individual search settings in advance.

The definition of default search settings or display options such as language selection or the amount of information displayed supports individualization and enhances efficiency. The majority of the termbases examined also allow users to perform a new search or modify their previous search on the results screen without returning to the search screen. This minimizes the users' memory load and supports user control.

The terminological entry is displayed in a way according to the default or personal settings or the display options selected on the results screen. In the terminological entry, some data fields such as definitions, sources or domain can be hidden. Thus, users get a better overview of the search results and can receive detailed information on a term by selecting this term from the result list or clicking a related icon or link. The majority of the analyzed termbases use icons sparingly in terminological entries. If icons are used, they have a tooltip with information that indicates the icon's meaning, e.g. icons used for feedback options or indicating the availability of additional information on a term in the result lists or terminological entry.

Help, Documentation and Error Messages. A termbase's web interface should be as intuitive as possible. However, help and documentation are necessary if users encounter problems in using the system or want to use advanced features. The documentation should be easily accessible and visible on the screen. It should be easy to search, focus on the user's task and list concrete steps (Nielsen 2010, 153). All but one of the studied termbases provide explanatory texts, FAQs, help sections and tooltips for search options, icons and other information on the user interface.

Systems should prevent errors from occurring in the first place. If errors occur, error messages should be written in clear language and avoid system codes. Error messages should be intelligible, state the problem and suggest solutions. In addition, systems should support users in recovering from errors, e.g. by undoing, editing or reissuing false commands without starting from scratch (Nielsen 2010, 142–45). The majority of the error messages that are displayed by the analyzed termbases as a result of an unsuccessful search inform users about the fact that their search term is not available in the termbase. Many error messages only indicate that no term has been found. More elaborate error messages also suggest solutions so that users can recover from this error, e.g. the systems suggest an alternative search term or prompt users to provide feedback. This increases the system's controllability and flexibility.

Individualization and Integration into other Platforms. Individualization of termbases means that users can tailor frequent actions with the termbase to their individual needs, e.g. searching in a certain language combination or domain. This enhances flexibility and efficiency in using the system. Three of the analyzed termbases enable users to save their default settings or display options. Here users can define the combination of (target) languages or the display of an entry's short or long version. In individual cases, they can also save their individual search history or an individual record online, or save and print an entry.

Some termbases also integrate or link to other tools. Two termbases allow users to search in (other) term collections. EuroTermBank allows users to import external data

and offers a Microsoft Word add-in. FAO TERM has a widget and the Microsoft Terminology Collection does not only return results from the termbase but also from the company's translation memory. Although three termbases provide a download of their content, none of their websites can be directly integrated into CAT tools.

Feedback Options and Social Features. Developments such as collaborative terminology work or crowdsourcing enable non-terminologists to contribute to terminology work with their input and feedback (Karsch 2015, 291; Kudashev 2013). The feedback options within the analyzed termbases include comments and term or change suggestions, i.e. user requests for adding terms or correcting the termbase's content. The primary means of communication with the terminology managers are forms or e-mails. The feedback forms require only a minimum amount of information from the users, e.g. contact details, comment and basic terminological information such as term or source. Some of the termbases studied facilitate sharing of an entry or entire page via e-mail or on social media with others.

3 Conclusion

The usability of termbases depends on their learnability, efficiency, memorability, prevention of errors and user satisfaction. It can be improved by adhering to principles that are applied to similar systems such as search engines, online dictionaries or online databases. Based on the users' previous experience with these systems, web interfaces of termbases should focus on the search function, i.e. the design of the search field and pre- and post-search filter options to refine the search. A simple search mode reduces the complexity of the system and enhances both learnability, i.e. novice users can immediately start to work with the system and memorability, i.e. casual users can quickly re-use the system after a period of non-use. Auto-complete or language suggestion functions can reduce the number of errors that might occur when users enter a search query. The satisfaction of the users depends on both a termbase's content, i.e. if users find the term and the information they want and the features they expect from a termbase, e.g. different search modes, filter and feedback options or sharing and printing of records. The efficiency is increased if users can achieve a high level of productivity with the system. However, further research should be undertaken to investigate the usefulness of the features and the usability of the analyzed termbases for various user groups.

References

1. Bank, Christina. 2012. "Die Usability von Online-Wörterbüchern und elektronischen Sprachportalen." *Information - Wissenschaft & Praxis* 63 (6). doi:10.1515/iwp-2012-0069.
2. Brinck, Tom, Darren Gergle, and Scott Wood. 2002. *Usability for the Web: Designing Web Sites That Work*. The Morgan Kaufmann series in interactive technologies. San Francisco: Morgan Kaufmann Publishers.

3. Cauna, Eduards. 2012. "Challenges and solutions for large multidomain terminology database." *Magyar Terminológia* 5 (1): 101–7. doi:10.1556/MaTerm.5.2012.1.9.
4. Chiocchetti, Elena, and Natascia Ralli. 2013. "Guidelines for collaborative legal/administrative terminology work". Bolzano: EURAC.
5. EuroTermBank Consortium. 2016. EuroTermBank Accessed April 01, 2016. <http://www.eurotermbank.com>.
6. Food and Agriculture Organization of the United Nations. 2016. FAO TERM PORTAL. Accessed April 01, 2016. <http://www.fao.org/faoterm>.
7. Go, Kentaro. 2009. "What Properties Make Scenarios Useful in Design for Usability?" In *Human Centered Design: First International Conference, HCD 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009 Proceedings*, edited by Masaaki Kurosu, 193–201. Berlin, Heidelberg: Springer.
8. Government of Canada. 2016. TERMIUM Plus® - the Government of Canada's terminology and linguistic term bank. Accessed April 01, 2016. <http://www.btb.termiuplus.gc.ca>.
9. Höge, Monika. 2002. *Towards a framework for the evaluation of translators' aids systems*. Helsinki: University of Helsinki.
10. ISO 26162:2012: *Systems to manage terminology, knowledge and content – Design, implementation and maintenance of terminology management systems*. Geneva: International Organization for Standardization.
11. ISO 9241-110:2006: *Ergonomics of human-system interaction - Part 110: Dialogue principles*. Geneva: International Organization for Standardization.
12. Karsch, Barbara I. 2015. "Terminology Work and Crowdsourcing." In *Handbook of Terminology*, edited by Hendrik Kockaert and Frieda Steurs, 289–303. Amsterdam: Benjamins.
13. Kudashev, Igor. 2013. "Quality Assurance in Terminology Management: Recommendations from the TermFactory project". Helsinki: Unigrafia.
14. Lemmetti, Mikko. 2001. "Usability of terminology management programs and databases: a survey study." Master's thesis, Department of English, University of Jyväskylä.
15. Marcos, Mari-Carmen et al. 2006. "Usability evaluation of online terminology databases." Accessed March 16, 2016. <http://www.upf.edu/hipertextnet/en/numero-4/usabilidad.html>.
16. Microsoft. 2016. Microsoft Terminology Collection. Accessed April 01, 2016. <https://www.microsoft.com/Language/en-US/Search.aspx>.
17. Nielsen, Jakob. 2010. *Usability engineering*. Amsterdam: Morgan Kaufmann.
18. Quesenbery, Whitney. 2001. "What Does Usability Mean: Looking Beyond 'Ease of Use'." Accessed April 01, 2016. <http://www.wqusability.com/articles/more-than-ease-of-use.html>.
19. Rubin, Jeffrey, and Dana Chisnell. 2008. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. 2nd ed. Indianapolis, IN: Wiley Pub.
20. SAP. 2016. SAPterm – SAP terminology database. Accessed April 01, 2016. <http://www.sapterm.com>.
21. Sevriens, A.P.M. 2010. "Improving usability for a terminology search website." Master thesis, Communicatie en Informatie, Tilburg University.
22. Translation Centre for the Bodies of the European Union in Luxembourg. 2016. IATE - InterActive Terminology for Europe. Accessed April 01, 2016. <http://iate.europa.eu>.
23. Tuominen, Alma I. 2012. "Investigating Usability: A Case Study of Wordfast Professional." Accessed April 01, 2016. <http://urn.fi/urn:nbn:fi:uef-20120668>.
24. United Nations. 2016. UNTERM – The United Nations Terminology Database. Accessed April 01, 2016. <https://unterm.un.org>.
25. World Trade Organization. 2016. WTOTERM. Accessed April 01, 2016. wtoterm.wto.org.

A Bilingual KRC Concordancer for Assisted Translation Revision based on Specialized Comparable Corpora

Firas Hmida, Emmanuel Morin, Béatrice Daille, and Emmanuel Planas

LINA-UMR CNRS 6241
University of Nantes, France
`firstname.lastname@univ-nantes.fr`

Abstract. Terminology is used all through the process of specialized translation. Indeed, many translators confirm that an error on terminology has a major impact on their work. Thus, a revision phase is necessary to validate the initial translation proposed by the translator. This paper deals with the assisted terminological revision in specialized translation from English to French. We propose a new generation of bilingual concordancers that takes as input a term and its translation, and provides not parallel but aligned Knowledge-Rich Contexts from specialized comparable corpora. Both the manual evaluation and a real experiment with student revisers show that our concordancer actually assists revisers despite the difficulty of the task.

Keywords: bilingual concordancer, Knowledge-Rich Contexts, specialized comparable corpus, collocations, revision, human translation

1 Introduction

In a survey conducted in Morin-Hernandez (2009, p. 143), 90% of the French translation professionals respond that an error on terminology has a major impact on translation work. Terminology is indeed crucial all through the translation process. Gouadec (2002) identifies three main steps in the translation process of specialized texts (specialized translation): pre-translation, translation and post-translation. The translation phase is itself divided into two classical sub-tasks conducted by translators: a translation task and a revision task.

Robert (2012, p. 95) identifies two main types of revision: the bilingual revision where the reviser carefully compares the original text (the source text written in the source language) and its translation (the target text written in the target language); and the monolingual revision where the translation is only revised in the target text. Both revisions can be conducted by the translator himself in a quest of a better production; or by a different translator called the reviser. The translation industry standards (German DIN 2345, European EN 15038, ISO 17100) imply the obligation for professional translators to review every translation by a third party translator or reviser.

In this paper, we are specifically concentrating on the bilingual revision where the reviser has to check different aspects of the first specialized translation draft (Delisle et al., 1999, p. 71). Thus, terminology comes as an important factor. To concretely illustrate the point, let us consider the translation of the term *blob* in the following text: *When the basalt magma first breaks out at the surface, the dissolved gases bubble off vigorously enough to carry **blobs** of magma into the air with them. The **blobs** may rise up 2,000 feet or more.*

Here, the translation of the term *blob* into French, in the field of volcanology, is not obvious. While in the general language the common translation of *blob* is *goutte* (drop, a scrap of something), a more suited translation is *projection* (spatter, splash). In this case, it is essential for the reviser to get access to textual contexts containing typical neighborhoods or providing useful information about the links between the terms involved in this translation (*blob* in the source language, and the translator’s choice in the target language, either *projection* or *goutte*) and the other terms and expressions of the field. These contexts are defined as Knowledge-Rich Contexts (KRCs) (Meyer, 2001). In the Cristal¹ project, we concluded with a list of attested KRCs that we automatically extract from prepared comparable corpora².

In this work, we aim at assisting revisers in the bilingual revision task by providing them with KRCs that will help them confirm or disapprove the translation that was already proposed. We will more precisely provide revisers with both source (EN) and target (FR) KRCs extracted from specialized comparable corpora, in a new generation of bilingual concordancer that we call KRCTool. We will show that this tool actually helps revisers in the framework of specialized revision.

2 Framework

We define at first the KRC concept, then we present the issue of classical bilingual concordancers in a revision framework.

2.1 Knowledge-Rich Contexts

Meyer (2001) introduces the notion of Knowledge-Rich Context to describe contexts that contain terms and relations between them in a specialized domain. These relations are usually expressed with lexical and syntactic patterns (Morin, 1999). For example, *An impact crater is caused by two celestial bodies impacting each other* is a KRC of the term *impact crater*, in which *is caused by* is a pattern reflecting a causality relation between *impact crater* and *celestial bodies*. All of these terms are from the domain of volcanology. KRCs have historically been introduced in the framework of terminology and knowledge extraction purposes.

¹ <http://www.agence-nationale-recherche.fr/?Projet=ANR-12-CORD-0020>

² Corpora that contain multilingual documents that are not translations of but share characteristics such as period and theme (Bowker and Pearson, 2002).

We consider that this notion refers also to other types of contexts, like the “examples” of Kilgariff et al. (2008). These examples are contexts identified thanks to collocations extracted from a general monolingual corpus. A Collocation is a regular co-occurrence of two items (base, collocate) within a specified field (Sinclair et al., 1970). A good command of collocations is an essential component of the proficiency of any language or specific discourse. Indeed, it is more correct to say *to prescribe medication* than *to write medication* in medical domain, or *to gush lava* instead of *to push lava* in volcanology. In these examples, *medication* and *lava* are the bases. Based on collocations, Kilgariff’s examples are undoubtedly considered as rich of knowledge since they illustrate typical neighborhoods in contexts. This knowledge is well appreciated by revisers. Planas et al. (2014) already showed that KRCs, based on collocations or relations between terms, can be useful to illustrate terms in specialized domain. Thus, we focus here on KRCs containing collocations of the source term or its proposed translation.

2.2 Bilingual Concordancers

Bilingual concordancers are resources more and more used to assist translators in terminological translation tasks. They often rely on parallel corpora. These tools allow translators to enter one term and, if this term occurs in the bilingual parallel corpus, to look how it was dealt with across the different contexts the tool returns. Perhaps one of the more popular concordancer among translators is the online service Linguee³, actually built from aligned parallel corpora.

In bilingual revision, the reviser who uses these kind of tools, that take only one term as input, has to enter the source or target terms independently. The link between source term (resp. the target term) and the term used as translation comes from the fact that contexts sentences are aligned in the parallel corpora. Despite their general usefulness, the main problem of classic concordancers is the scarcity of parallel corpora, especially in specialized domain. Furthermore, contexts proposed by Linguee are generally quite broad and lack specific knowledge that could be found in specialized corpora. A special use of SketchEngine⁴, the “bilingual word-sketch”, allows the input of the couple (source term, target term) and provides a series of available collocations from which some context can be retrieved. These use large corpora (parallel and comparable) in general domain, and different alignment schemata where the compositional term alignment is used (Baisa et al., 2014). The multilingual sentence alignment from comparable corpora drew much research attention. Rauf and Schwenk (2011) shows that parallel sentences are quite scarce in comparable corpora, especially in specialized domain.

In this paper, we rely on the comparability of comparable corpora collected from specialized texts. We build a bilingual concordancer called KRCTool that provides not parallel but aligned KRCs to help in revising a pair (source term/proposed translation) given as input.

³ <http://www.linguee.fr/>

⁴ <https://www.sketchengine.co.uk/bilingual-word-sketch/>

3 Method

In the comparable corpora, parallel or lexically similar contexts are rare. It would be even more restricted to align KRCs on the base of their lexicon. Consequently, our aim is to determine bilingual “properties” which enable the reviser to build transition bridges between source and target contexts. We then allocate to each source KRC an equivalent target KRC based on these properties. We propose a methodology first based on **extraction of KRCs**: for each (source term/proposed translation), we extract the collocations of the source term and its proposed translation and then retain the sentences that contain the automatically translated collocates. These sentences are considered as KRCs. And after based on **alignment of KRCs**: the bilingual sentences resulting from the previous step will be filtered and aligned.

3.1 KRC Extraction

Mammino (1995) approached the issue of specialized terms and their use, that are faced by translators without in-depth knowledge of the terminology. In this case, a translation that does not respect the standard collocations of the domain may be negatively perceived by revisers (Musacchio and Palumbo, 2008). Revisers frequently look for approximations of the source collocation, in the target language. If the literal translation is correct, it would be unwise to try at all costs to avoid it, because it may allow referential and pragmatic equivalences (Newmark, 1988, p. 68-96). Here, our purpose is not to translate collocations, but to provide relatively close collocations, that can help revisers check if the proposed translation is in its typical context.

First, we implement the z-score to automatically extract collocations according to their syntactic structures: (T, Adj), (T, N) and (T, V), with T the single term we want to illustrate. Then, we align collocations, pairing collocates belonging to the same grammatical category. Even if the overlap between collocations and multi-word terms is a well-known problem in collocation extraction, here, we do not distinguish between these two phenomena that may share co-occurrence and syntactic criteria.

3.2 KRC Alignment

The obtained KRCs at this stage are aligned only on the basis of collocations, which often prove to be insufficient. Therefore, we will refine the KRC alignment using other anchor points in addition to collocations. Our goal now is to filter and align them:

1. **filtering criteria:**

- **context length:** short sentences could not contain more knowledge than the collocation. Conversely, it is very difficult to consult sentences that are very long, also they may illustrate irrelevant information for the revision. As Kilgarrriff et al. (2008) we retain only sentences containing between 10 and 20 full words.

- **pronouns:** Kilgarriff et al. (2008) penalize contexts that contain pronominal anaphora, since it may refer to text unities in previous sentences. We assume that pronouns inside contexts are less problematic because they can refer to unities in the same sentence. We eliminate only contexts starting with a pronoun.
- **affirmative contexts:** Kilgarriff et al. (2008) prefer affirmative sentences rather than interrogative ones. We also retain this criterion to filter out interrogative contexts.
- **context complexity:** this criterion was also addressed by Didakowski et al. (2012) to measure the readability of the sentence. We follow the same strategy using a dependency parser to filter complex contexts. In our case, we use the sum of the scores of all possible parse trees for a given sentence to measure the complexity: the more complex the context is, the greater is the sum of all its possible trees.

2. alignment criteria:

- **number of cognates:** we consider cognates as two words starting with the same 4 characters as Léon (2008). They represent transition bridges easily detected by the reader, in pairs of source and target contexts. Contexts sharing at least one cognate, will be aligned.
- **number of translated simple terms:** despite their scarcity in the corpus, sentences containing translated terms are exceptionally operational for the reviewer. The single word terms of the studied corpus were extracted by a dedicated terminological tool. Contexts containing at least one simple term and its translation will be aligned.

4 Manual Evaluation

To evaluate the quality of the aligned KRCs, we manually prepared reference KRCs for each studied term and its translation. In this section, we present the used corpora, the reference data and the experiments.

4.1 Corpora and Bilingual Dictionary

This evaluation was carried out on specialized comparable corpus built by Josselin-Leray (2005) and obtained through a thematic research from newspapers and magazines in the field of volcanology. This corpus is composed of English and French scientific documents containing roughly 400,000 words per language. They have been cleaned and standardized through TermSuite⁵ that also extracts terminology. For the automatic alignment of collocations, we used ELRA⁶, a bilingual dictionary of general language (EN-FR) containing 145,542 entries. It also contains the POSs of entries.

⁵ <https://logiciels.lina.univ-nantes.fr/redmine/projects/termsuite>

⁶ http://catalog.elra.info/product_info.php?products_id=666

4.2 Evaluation Data

Bilingual Aligned KRCs were manually prepared for 15 pairs of single-word terms essential for the volcanology domain. Here are some examples: *basalt/basalte*, *cinder/scorie*, *volcan/volcan*, *eruption/éruption*... The multi-word terms have been excluded for the reason that the identification of complex terms collocations can be treated as a separate issue that we do not regard in this work. The process that we followed to prepare the reference KRCs was:

1. For each pair of terms, we manually identify the source and target collocations in which collocates are translations. Then, we extract contexts that contain these collocations. Here, experts were solicited to check the manual translation.
2. We checked manually if contexts provided for each collocation were valid. A context is valid only if the collocation in question is valid within it.

4.3 Experimentation

We applied our method on the 15 pair of terms and we evaluate the bilingual KRCs aligned with and without filters. The aligned KRC pairs were manually validated if at least one of the following conditions is valid:

1. the alignment criteria are also valid within a window of 7 words (approximately) containing the term in question or its proposed translation. For example:
 - pair of translation: *lava*, *lave*
 - aligned collocations: (*lava*, *basaltic*) and (*lave*, *basaltique*)
 - source KRC : *Shield cones are broad, slightly domed volcanoes built primarily of fluid, **basaltic lava**.*
 - target KRC: *Volcan bouclier, volcan de forme ovale, très aplati, dû à l'accumulation de coulées de **lave basaltique** fluide.*

Here, the concentration of the alignment criteria within a window of words that can be easily consulted, help to validate the pair of the aligned KRCs.

2. the “global topics” of the two KRCs are similar. The alignment criteria, which are mainly lexical, could be non relevant towards the reviewer. In this case, if the topics of the contexts in question are similar, they can be considered as a bridge transition between the contexts. In the following example, KRCs have been validated thanks to the similarity of the subjects that they treat:
 - pair of translation: *cinder*, *scorie*
 - aligned collocations: (*cinder*, *incandescent*) and (*scorie*, *incandescent*)
 - source KRC: *Strombolian eruptions are named for Stromboli volcano off the west coast of Italy, where a typical eruption consist of the rhythmic ejection of **incandescent cinder**, lapilli, and bombs to heights of a few tens or hundreds of feet meters.*
 - target KRC: *Le dynamisme strombolien s'exprime par des explosions rythmiques qui projettent des blocs et des **scories incandescentes**.*

Table 1. Evaluation of aligned KRCs: with and without filters

Corpora	# terms	# aligned terms	# pairs aligned coll.	# contexts coll.	# of KRCs	# pairs aligned of KRCs	P. valid pairs aligned
without filters							
Vulcano EN 15	10		23	677	309		43,04%
Vulcano FR 14				665			
with filters							
Vulcano EN 15	10		16	241	157		61%
Vulcano FR 14				296			

4.4 Results

The analysis of table 1 shows that the aligned collocations are productive: each collocation pair produces on average 28 contexts without filter, and 15 with filter, for each language. We note that even if the application of filters deteriorate the number of aligned KRCs, it significantly improves the precision of the alignment criteria since it moves from 43% to 61%. We could not provide bilingual contexts for five pairs of terms. Some of these pairs have a too small number of extracted collocations or only one syntactic structure. For the others, the alignment method act as a filter and eliminates contexts in both languages.

5 Experiment with Revisers

After having studied the quality of bilingual KRCs, we perform real experiment with student revisers using the KRCTool.

5.1 Experimental Data

We had conducted an experiment in a previous framework where 11 second year Master students translated the same text from English to French. For our current experiment, we used the same English text, and selected one of the student translations for the revision task. We retain one of the most perfectible ones. We identified three terms in the source text; and changed the translation of these items with more common translations in the target text. We then expected the revisers to correct these “lazy” translations by terms more specific to the domain of volcanology, with the use of the KRCTool. Table 2 contains the source and the translation terms that we changed, with acceptable translations.

Table 2. Source and changed terms

source term	modified translation	correct translations
cinder	débris	scorie, cendre
vesicle	poche	vacuole, vésicule
blob	boule	paquet, projection

Here is a detailed view of the reasoning we expected. When the couple *blob* and *goutte* is searched in KRCTool, only one target KRC containing *goutte* is shown. Nevertheless, this KRC shows a use of *goutte* which is restricted to an in-vitro experiment, that does not fit with *blob* here. A good reviser should here disapprove *goutte* and search for an alternative solution.

Instead, if *blob* and *projection* are searched in the KRCTool, as suggested by the available translation, instances of *projection de lave* are displayed along with *blob of magma*. This provides a more acceptable translation for *blob*.

5.2 Protocol

In order to test whether the KRCTool would help the revisers or not, we designed the following protocol. We had two groups A and B of first year students from a Master in Professional Translation. We divided each group into two sub-groups and asked each sub-group to work on a different part of the text, as sums-up table 3. This was done to prevent and smoothen any specificity of these text parts that may influence the revision task. In a first phase, students A had to revise the translation text with their usual resources like Linguee, Le Grand Dictionnaire Terminologique or CRISCO (synonyms): the objective was to correct as best as possible the text so as to get a good translation. In a second phase, the same students A had to correct the translated text only using KRCTool. Students B did the same task, but started in Phase 1 with the use of the KRCTool first. In Phase 2, they made use of their usual resources.

Table 3. Group repartition

Group A	Text 1	Text 2	Time (min)
Phase 1:common res.	Aa	Ab	20
Phase2:KRCTool	Ab	Aa	20
Group B	Text 1	Text 2	
Phase 1:KRCTool	Ba	Bb	20
Phase2:common res.	Bb	Ba	20

5.3 Results

Based on table 4, KRCTool proved to be useful for correcting the translation of the three terms. For each term, a revised translation was provided by 1 to 4 students (out of 14) with the use of KRCTool. All revised translation were correct. In an post survey, students declared that the KRCTool provided them with specific and specialized contexts that they did not find in their usual resources. We see that group B provided more corrections using the KRCTool than group A. We believe this is because group B started in Phase 1 by using the KRCTool. Whereas Group A first used common resources in Phase 1, and then used the KRCTool only in Phase 2: hence, most of the terminology searches for group

Table 4. Revision results. (Nb: number of performed revision; x: performed revision; possible translations provided by KRCTool for *cinder*: *scorie*, *cencre*, *débris*; for *blob*: *projection*, *paquet* and *boule*; for *vesicle*: *vésicule*, *vacuole* and *poche*; for *bubble off*: *partent*; and for *spewed out*: *sort*).

Term	Nb	Aa1	Aa2	Ab6	Ab7	Ab8	Ba1	Ba2	Ba3	Ba4	Bb6	Bb7	Bb8	Bb9	Bb10
With Common resources															
cinder	6	-	-	-	-	x	-	x	x	-	-	x	x	x	-
blobs	3	-	-	-	-	-	-	-	x	x	-	-	-	x	-
vesicles	4	x	x	-	-	-	-	x	x	-	-	-	-	-	-
Total (T1)		1	1	-	-	1	-	2	3	1	-	1	1	2	-
With KRCTool															
cinder	2	-	-	-	-	x	-	x	-	-	-	-	-	-	-
blobs	4	-	-	-	-	-	-	-	x	-	-	x	x	x	-
vesicles	1	-	-	-	-	-	-	-	-	-	-	x	-	-	-
Total (T2)		-	-	-	-	1	-	1	-	1	-	2	1	1	-
T2 ≥ T1 ≥ 1		-	-	-	-	x	-	-	-	x	-	x	x	-	-
1 < T1 < T2		x	x	-	-	-	-	x	x	-	-	-	-	x	-

A were processed in Phase 1 using common resources: there was less searches left for KRCTool. Two students (Ba1 and Bb7) provided more corrections with the KRCTool than with other common resources. Table 4 also shows that using the KRCTool, four students among the 13 ones which carried out corrections have successfully accomplished the same revision as with common resources, or better. However, five students performed a better revision based on common tools. We have to admit that these students were only first year Master and did not have previous knowledge of this specialized domain to correct all the terms as a professional reviser would. In average, students provided more corrections with common resources that provide more output. Debutant students tend to be seduced by the quantity rather than the quality of the resources.

6 Conclusion

This paper proposes KRCTool as an example of a new generation of bilingual concordancers that takes as input a source and a target term and provides aligned KRCs from specialized comparable corpora, for an assisted revision purpose. KRCTool is based on a methodology that uses collocations, cognates and the translation of simple terms as anchor points for the identification and the alignment of KRCs in specialized comparable corpora. The manual evaluation shows that the KRCs we obtain are quite acceptable for a manual revision. The experiment performed with revisers confirms indeed that KRCs proposed by the KRCTool actually assist revisers in a translation revision task. The study we carried out deals with qualitative aspects of the obtained KRCs that we wanted to completely control. That is why our experiments relied on few terms. Further experiment should be driven for confirming our findings.

Bibliography

- Baisa, V., M. Jakubek, A. Kilgarrieff, V. Kov, and P. Rychl (2014). Bilingual word sketches: the translate button. In *EURALEX*, Bolzano, Italy, pp. 505–513.
- Bowker, L. and J. Pearson (2002). *Working with specialized language: a practical guide to using corpora*. Routledge.
- Delisle, J., H. Lee-Jahnke, and M. C. Cormier (1999). *Terminologie de la Traduction: Translation Terminology*. John Benjamins Publishing.
- Didakowski, J., L. Lemnitzer, and A. Geyken (2012). Automatic example sentence extraction for a contemporary German dictionary. In *EURALEX*, Oslo, Norway, pp. 343–349.
- Gouadec, D. (2002). *Profession: traducteur*. La Maison du dictionnaire.
- Josselin-Leray, A. (2005). *Place et rôle des terminologies dans les dictionnaires généraux unilingues et bilingues: étude d'un domaine de spécialité: volcanologie*. Ph. D. thesis, Université de Lyon 2.
- Kilgarrieff, A., P. Rychlý, M. Husák, M. Rundell, and K. McAdam (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *EURALEX*, Barcelona, Spain, pp. 425–432.
- Léon, S. (2008). *Acquisition automatique de traductions d'unités lexicales complexes à partir du Web*. Ph. D. thesis, Université de Provence.
- Mammìno, L. (1995). *Il linguaggio e la scienza*. Torino: Società Editrice Internazionale.
- Meyer, I. (2001). Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In D. Bourigault, C. Jacquemin, and M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology*, pp. 279–302. John Benjamins Publishing Company.
- Morin, E. (1999). Using lexico-syntactic patterns to extract semantic relations between terms from technical corpus. In *TKE*, Innsbruck, pp. 268–278.
- Morin-Hernandez, K. (2009). *Revision as a key function of translation quality management in a professional context*. Ph. D. thesis, Université Rennes 2.
- Musacchio, M. T. and G. Palumbo (2008). Shades of Grey: A Corpus-driven Analysis of LSP Phraseology for Translation Purposes. In *Corpora for University Language Teachers*, pp. 69–79. Bern: Peter Lang.
- Newmark, P. (1988). *A Textbook of Translation*. Prentice-Hall International.
- Planas, E., A. Picton, and A. Josselin-Leray (2014). Exploring the Use and Usefulness of KRCs in Translation: Towards a Protocol. In *TKE*, Berlin, Germany, pp. 188–228.
- Rauf, S. A. and H. Schwenk (2011). Parallel sentence generation from comparable corpora for improved smt. *Machine translation* 25(4), 341–375.
- Robert, I. S. (2012). *La révision en traduction: les procédures de révision et leur impact sur le produit et le processus de révision*. Ph. D. thesis, University of Antwerp.
- Sinclair, J. M., S. Jones, and R. Daley (1970). *English Lexical Studies. Final Report of O.S.T.I. Programme C/LP/08*. Department of English.

Semi-automatic Evaluation of Terminological Web-crawled Corpora

Lotte Weilgaard Christensen

University of Southern Denmark, Kolding, Denmark

lotte@sdu.dk

Abstract. This paper presents a method for evaluating the suitability of web-crawled corpora for terminological analyses. Since the contents of web-crawled corpora are unknown, the need arises for testing whether such corpora comprise a sufficient quantity of terminological information, the focus being on linguistic and conceptual coverage. The analyses are based on the corpus management system Sketch Engine, combined with knowledge patterns based on the valency of Danish verbs. The results originate from corpora established for exam purposes and used by students. So far Sketch Engine has been used primarily by lexicographers. However, the paper demonstrates that terminologists may use the program for far more than term extraction.

Keywords: corpus evaluation; web-crawled corpora; corpus tools; Sketch Engine; knowledge patterns; terminology extraction; knowledge extraction

1 Introduction

This paper aims to discuss methods for evaluating to what extent a web-crawled corpus compiled from the Internet comprises information of relevance for terminology work. The challenge to be faced by users of such a corpus is the fact that they do not know its contents, and that consequently, there is a risk of performing terminology work on an unknown basis. In the article ‘Getting to know your corpus’, aimed at lexicographical investigations, Adam Kilgarriff [6] raises some questions of great relevance in this connection: “But can we trust a crawled corpus?”, and “How do we know what is in it, or if it does a good job of representing the language?”

However, knowing if the amount of knowledge represented by the linguistic data of a given corpus is sufficient for our investigation is necessary for terminological investigations. Even if a domain corpus comprises a large number of terms, it will not necessarily include sufficient terminological information to identify the semantic relations required to establish conceptual systems, nor will it necessarily comprise linguistic data suitable as input for definitions.

The web-crawled corpora will be tested using the corpus management system Sketch Engine, widely used by lexicographers [5]. To my knowledge, in connection

with terminological investigations, Sketch Engine has primarily been used for extracting term candidates [5], whereas little attention has been given to the extraction of other terminological information by means of the system. The article will demonstrate that Sketch Engine is far more capable of supporting terminology work than what has been described until now. Moreover, the Sketch Engine team is in the process of developing methods for automatic extraction of hierarchical relations as well as definitions [1]. Thus, the aim of the article is to discuss how web-crawled corpora compiled by Sketch Engine may be tested, and to demonstrate how Sketch Engine may also support the retrieval of terminological information at the conceptual level. In the semi-automatic evaluation methodology, a subset of Danish knowledge patterns have been implemented which are suitable for retrieval of knowledge-rich contexts (KRC). Knowledge-rich context has been defined by Meyer [7] as “a context indicating at least one item of domain knowledge that could be useful for conceptual analysis”.

The retrieval of information from corpora and the evaluation of corpora for terminological purposes are to some extent two sides of the same coin. In this paper, focus will be on monolingual information retrieval. Besides, the approaches needed for different languages, depending on the corpora and the corpus analysis functions available will be compared. In that connection, the importance of finding simple methods easily applied by all user groups must be emphasized.

The rest of the paper is organized as follows: in Section 2, I describe the background, in Section 3, the criteria to be used for corpus design are discussed. Section 4 comprises a general description of Sketch Engine. Sections 5 and 6 deal with issues of linguistic and conceptual coverage, including Sketch Engine functions and knowledge patterns. Section 7 ends with some concluding remarks.

2 Background

This investigation focuses on specialized corpora used for exam assignments in terminology courses in which students are expected to demonstrate their mastering of the methodology of terminology. In the typical assignment, students will be asked to construct a conceptual system comprising 10 to 20 concepts and to write definitions of some of the concepts in question.

Compiling web-crawled corpora for the above purposes revealed that although they comprised a large number of term candidates, even corpora consisting of more than 50,000 tokens might not necessarily include sufficient amounts of elements of knowledge to enable students to carry out the terminology tasks required. And when students have been given a corpus for terminological investigations for the purposes of an exam, they naturally expect the corpus to be suitable for the retrieval of conceptual information and not only for the identification of terms.

3 Criteria of Corpus Design

Obviously, the work load involved in building a web-crawled corpus is smaller than the one needed to compile a well-designed corpus. However, if the corpus texts used for a terminology project turn out to be of inferior quality, the end result of the project will likewise be of inferior quality and turn out a very expensive one. Thus, since the contents of web-crawled corpora are not known, methods of evaluating such corpora must be found.

Bowker and Pearson [3] define corpus as 'a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria'. Indeed, it is generally agreed in terminology literature that well-designed corpora should be compiled according to specific criteria determined by the goals of the task in hand, criteria such as text type and function, reliability, level of expertise, and domain coverage, including both linguistic and conceptual coverage.

In what follows, a method of semi-automatic validation of corpora, focusing on the criteria of linguistic and conceptual coverage will be presented. Those criteria have been chosen because a domain-specific corpus must find itself at a level of linguistic and conceptual coverage that will suffice for the purpose defined; if not, the students will not be able to use it to demonstrate their ability to apply terminological methodology. Here the term linguistic coverage refers to a sufficient number of terms in a corpus. Likewise, the term conceptual coverage refers to a sufficient amount of knowledge-rich contexts for the purpose of a given exam assignment. As indicated earlier, the testing is based on Sketch Engine combined with Danish knowledge patterns.

4 The Corpus Management System Sketch Engine

Sketch Engine is an advanced commercial corpus management system. It primarily distinguishes itself from other corpus tools by comprising an integrated piece of software called WebBootCat, compiling texts into a corpus by crawling the web. In order to create a corpus from the Web, the user is asked to specify 3 to 20 seed words, i.e. key words or multi-word expressions, from the subject domain to be investigated [5]. In a sense, at this stage the seed words are the criteria defining your corpus.

In addition to basic functionalities known from other corpus analysis tools, the corpus analysis tool of Sketch Engine offers additional advanced functionalities, some of which will be described below.

Moreover, Sketch Engine includes large LGP corpora in sixty languages [6], to be applied as reference corpora when extracting lists of term candidates. For English, to name an example, the British National Corpus is available [5]. For Danish, large LGP corpora have also been added in recent years.

In order to apply the advanced functionalities of Sketch Engine, so-called 'high-

level resources' must be integrated, including: a tokeniser, a lemmatizer, a part-of-speech tagger, and a parser or 'sketch grammar' [5]. A sketch grammar identifies possible relations of words to a keyword [8].

Sketch Engine does not support all languages with high-level resources. This means that depending on the language used, there will be substantial differences as to what analyses can be carried out using Sketch Engine, as well as to the ways in which it will support the terminological investigations. Since Danish is one of the languages for which high-level resources are not available, at least not for the untagged domain specific user corpora, users must rely on functions based on statistical calculations and find pragmatic approaches to information retrieval as well as to testing the usability of their corpora.

5 Linguistic Coverage

Below, it will be illustrated how the linguistic coverage of a corpus can be evaluated using Sketch Engine. The examples on *bicycles* originate from a corpus compiled for an exam assignment on this topic.

5.1 Term Extraction

The first step when testing the linguistic coverage of a corpus is to apply the term extractor function offered by Sketch Engine. This function compares the domain specific corpus to a reference corpus. The term extractor generates a file consisting of two columns called 'Single-word' (in earlier versions 'keywords') and 'Multi-word' (in earlier versions 'terms'), respectively. From a terminological perspective, the new designations are more motivated since both columns represent term candidates. The columns are shown in Fig. 1 below, retrieved from an English corpus on *bicycles*, compiled for this purpose, totaling 73,961 tokens. From the frequency information in the columns, a concordance list can be accessed directly.

Bicycles: Extracted keywords / terms ?

[Change extraction options](#) Download singlewords: [TBX CSV](#). Download multiwords: [TBX CSV](#).

Singlewords and multiwords are ordered by [keyness score](#). The score and corpus frequency (leading to the respective concordance) are displayed in parentheses. Highlighted words were used as seeds in a previous WebBootCaT run within this corpus.

[<< Back to corpus files](#)

Use WebBootCaT with selected words

Single-word	Score	F	RefF	Multi-word	Score	F	RefF
<input type="checkbox"/> bicycles	W 649.03	277	49,603	<input type="checkbox"/> diamond frame	W 628.85	56	111
<input type="checkbox"/> wsd	W 461.52	43	715	<input type="checkbox"/> top tube	W 495.33	49	1,563
<input type="checkbox"/> gazelle	W 435.97	60	7,232	<input type="checkbox"/> electric bicycle	W 470.08	45	1,091
<input type="checkbox"/> mixte	W 410.57	37	271	<input type="checkbox"/> human power	W 300.92	30	1,697
<input type="checkbox"/> bicycle	W 390.50	455	157,834	<input type="checkbox"/> road bike	W 293.16	41	7,586
<input type="checkbox"/> bikes	W 371.60	478	175,593	<input type="checkbox"/> electric motor	W 270.89	56	17,382
<input type="checkbox"/> sportive	W 339.13	37	3,060	<input type="checkbox"/> weight limit	W 215.21	24	3,446
<input type="checkbox"/> batavus	W 326.58	29	86	<input type="checkbox"/> privacy invasion	W 198.28	18	405
<input type="checkbox"/> handlebars	W 318.99	55	12,342	<input type="checkbox"/> electric bike	W 192.89	20	2,308
<input type="checkbox"/> motorized	W 318.68	81	24,325	<input type="checkbox"/> coaster brake	W 189.71	17	239
<input type="checkbox"/> schwinn	W 294.27	43	8,492	<input type="checkbox"/> bicycle attention	W 170.52	15	1
<input type="checkbox"/> ebike	W 275.98	25	355	<input type="checkbox"/> sportive bicycle attention	W 170.52	15	0
<input type="checkbox"/> zonar	W 248.24	22	73	<input type="checkbox"/> sportive bicycle	W 170.52	15	0
<input type="checkbox"/> moped	W 238.86	32	6,721	<input type="checkbox"/> good quality chain	W 170.52	15	7
<input type="checkbox"/> pedals	W 228.73	68	30,661	<input type="checkbox"/> quality chain	W 169.50	15	80
<input type="checkbox"/> bike	W 222.29	940	606,870	<input type="checkbox"/> brake horsepower	W 167.34	15	253
<input type="checkbox"/> hollandbikeshop	W 215.73	19	0	<input type="checkbox"/> adult content	W 163.93	18	3,209
<input type="checkbox"/> mopeds	W 205.52	23	3,497	<input type="checkbox"/> level ground	W 155.14	16	2,239
<input type="checkbox"/> pedego	W 204.06	18	23	<input type="checkbox"/> mountain bike	W 148.74	37	23,581
<input type="checkbox"/> shimano	W 203.75	33	10,833	<input type="checkbox"/> power-assisted bicycle	W 147.62	13	29
<input type="checkbox"/> pedelecs	W 177.18	16	340	<input type="checkbox"/> bike shop	W 144.78	18	5,347
<input type="checkbox"/> crossbar	W 171.04	21	5,102	<input type="checkbox"/> aluminum frame	W 140.46	15	2,782
<input type="checkbox"/> ebikes	W 168.00	15	195	<input type="checkbox"/> maximum speed	W 137.58	19	7,370

Fig. 1. Term extraction in Sketch Engine (not complete)

For languages not supported by the high-level resources, the term extractor function only generates a list of single-words, i.e. only single-word term candidates may be found. This means that Danish users must apply a more pragmatic approach in order to extract multi-word terms, using the concordance function. This approach has already been described in connection with other corpus analysis tools.

However, for languages in which many concepts are represented by composite terms, the next step in evaluating the degree of linguistic coverage is to enter a generic term, e.g. in our case *cykel* as a common head, which is the Danish word for *bicy-*

cle, and to search on this term as a truncated character string in order to identify potential types of the generic term which may indicate subordinate concepts. This search result may be re-sorted alphabetically using the search node so that all instances of the same composite terms are grouped together. On the basis of the concordance list sorted by node form, it is possible to generate a frequency list of the node forms in question. In this way, the search result from the concordance list can be narrowed down, making it easier to use than a multi-page concordance list, as illustrated in Fig. 2:

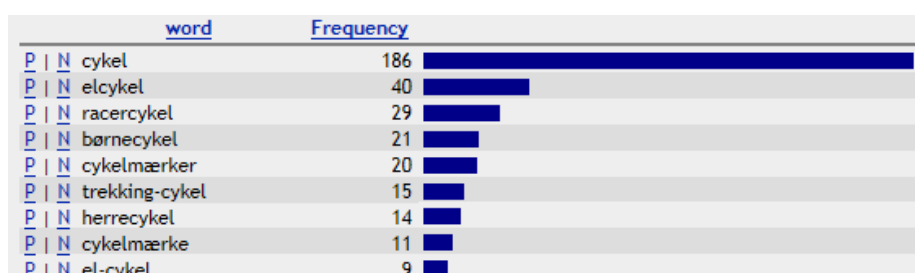


Fig. 2. Extract of frequency list of composite term candidates including *cykel* as generic term

The next step will be to search for the generic term without truncation, and to sort the search result using the word strings immediately to the left hand-side and immediately to the right-hand side of the generic term, respectively, in order to identify recurrent patterns that might represent multi-word terms, such as *elektrisk cykel* (*electric bicycle*).

5.2 Word Sketch Used for Term Extraction

Sketch Engine has its name from the word sketch which is a core function consisting of a one-page summary of a specific word's grammatical and collocational behaviour [5]. In other words, the word sketch is a list containing different recurrent patterns of the word searched for, according to the grammatical function of the word. The word sketch function requires a sketch grammar, cf. section 4. Fig. 3 shows a word sketch for the noun *bicycle*, retrieved from the English corpus on *bicycles*.

The columns labeled 'modifier' and 'modifies' offer information on term candidates. In the first case, *bicycle* is the head of potential multi-word terms with modifiers such as *electric*, *electric-assisted*, *power-assisted*, or *city*. In the second case *bicycle* is the modifier in nouns such as *bicycle shop*, *bicycle helmet*.

Compared to the manual work that must be carried out by the terminologist analyzing concordances, the word sketch provides him or her with a quick and easy overview of the recurrent patterns in which multi-word term candidates may occur.

Any word occurring as a frequent word together with the word searched for in word sketch will be provided with a frequency number. Via this number the relevant concordance list can be accessed directly.

bicycle (noun)
Bicycles freq = 699 (7,899.73 per million)

object_of 222 2.50	adj_subject_of 18 1.80	modifier 376 1.60	modifies 127 0.50	possessor 39 5.50
motorize <u>53</u> 12.52	online <u>3</u> 12.06	electric <u>99</u> 12.32	attention <u>15</u> 11.60	men <u>8</u> 12.11
assist <u>15</u> 10.91	short <u>4</u> 12.04	power-assisted <u>19</u> 10.60	shop <u>7</u> 10.30	women <u>7</u> 11.92
define <u>10</u> 10.34		electric-assisted <u>16</u> 10.38	path <u>6</u> 10.30	woman <u>12</u> 11.20
operate <u>7</u> 9.93		speed <u>12</u> 9.86	helmet <u>5</u> 10.12	man <u>7</u> 11.00
classify <u>6</u> 9.72		electric <u>11</u> 9.72	option <u>5</u> 10.09	lady <u>4</u> 11.00
sell <u>6</u> 9.66		city <u>9</u> 9.50	lane <u>5</u> 10.06	
design <u>6</u> 9.53		robust <u>8</u> 9.39	law <u>5</u> 9.69	pp_obj_as 29 9.70
be <u>21</u> 9.50		round <u>7</u> 9.23	operator <u>3</u> 9.48	class <u>6</u> 12.29
allow <u>5</u> 9.34		conventional <u>7</u> 9.20	market <u>3</u> 9.44	classify <u>3</u> 11.45
consider <u>5</u> 9.34		standard <u>7</u> 9.15		pp_with 24 2.90
propel <u>4</u> 9.07		pedal <u>6</u> 8.92	and/or 68 0.70	motor <u>11</u> 12.87
equip <u>4</u> 9.05		frame <u>6</u> 8.74	tricycle <u>8</u> 11.67	pp_obj_to 12 1.80
find <u>4</u> 8.91		gas <u>4</u> 8.43	vehicle <u>3</u> 10.17	apply <u>3</u> 12.12
mean <u>3</u> 8.64		trailz <u>4</u> 8.43	motorcycle <u>3</u> 10.10	refer <u>3</u> 11.83
use <u>4</u> 8.62		motorized <u>4</u> 8.42	wheel <u>3</u> 9.71	pp_if 8 15.00
include <u>3</u> 8.50		low-speed <u>4</u> 8.42	motor <u>3</u> 9.67	zonar <u>7</u> 13.90
ride <u>3</u> 8.13		ezip <u>4</u> 8.42	pp_obj_of 42 2.50	pp_at 7 3.30
		ordinary <u>4</u> 8.41	use <u>10</u> 12.34	hollandbikeshop <u>4</u> 13.54
subject_of 92 1.70		regular <u>4</u> 8.40		speed <u>3</u> 11.54
be <u>55</u> 10.50		lady <u>4</u> 8.33		
do <u>7</u> 10.29		lovely <u>3</u> 8.02		
appear <u>3</u> 9.97		suitable <u>3</u> 8.02		
use <u>3</u> 9.59		step-through <u>3</u> 7.95		
have <u>5</u> 8.89		diamond <u>3</u> 7.77		

Fig. 3. Word sketch for *bicycle*

For Danish, the word sketch function is not available for crawled user corpora. At present, for Danish or other languages without high-level resources, it is necessary to retrieve the information comprised by the word sketch by analyzing the concordance lists manually.

In his article ‘Getting to know your corpus’, Kilgarriff [6] argues that keyword lists, combined with a Sketch Engine function comparing two corpora, based on a model called simple math, is an essential support for the user wanting to gain an overview of the contents of a corpus, since in Kilgarriff’s words, a keyword list “takes frequency lists as summaries of the two corpora, and shows us the most contrasting items” [6]. This is true as far as linguistic coverage is concerned. However, this will not suffice for terminological investigations.

6 Conceptual Coverage

For terminological investigations, we obviously need a method to secure sufficient conceptual coverage as well. The next natural step will be to identify possible relations among concepts in order to be able to work out preliminary drafts of concept systems. The searches mentioned above for generic terms constituting the shared heads of composite terms or of multi-word terms will give you an impression, not just of the degree of linguistic coverage, but frequently also of potential terms representing subordinate concepts entering into type relations with the generic term (concept). However, for students to be able to carry out thorough terminological investigations, it must be secured that the corpus contains explicit knowledge-rich contexts.

6.1 Knowledge Patterns for Danish

Previously, I have analyzed domain specific corpora with the object of identifying knowledge patterns for Danish, my main focus being on recurrent patterns of verbs and their surroundings. My approach was originally based on a valency theory called the Pronominal Approach [4], building mainly on syntactic criteria. Many Danish verbs are formed analytically by means of e.g. prepositional objects and particles, as described in Weilgaard Christensen [9,10]. Thus, the approach in question enables identification of search patterns consisting of a verb together with a specific preposition. This character string approach makes it possible to eliminate terminologically irrelevant patterns (noise), and thus to narrow down the search result. As a natural consequence, an important insight achieved is that optional arguments which occur with prepositions become mandatory when they are used for the retrieval of terminological data [9,10].

In fact, for some verbs it is possible to predict with a considerable degree of certainty which terminological information can be identified using the knowledge patterns. The degree of predictability is particularly high for verbs identifying concept-related information, especially relations among concepts. For other patterns, the degree of predictability is somewhat smaller since they result in rather different terminological information or noise.

Therefore, I have introduced the concepts of strong and weak knowledge patterns, respectively [9,10]. ‘Strong knowledge patterns’ are patterns with a high degree of proportionality or even constant relations of proportionality with the categories of terminological information. Table 1 shows some important strong knowledge patterns of Danish verbs. One example of a constant relation is the Danish verb *inddele* (*subdivide*) together with the preposition *i* (*into*). In this case, the verb phrase will always identify a superordinate concept followed by subordinate concepts as objects in the prepositional construction, as shown in example (1). On the basis of this example, a small concept system can be sketched.

Table 1. Important strong knowledge patterns based on Danish verbs

Terminological information category identified	Verb + preposition
superordinate concept + subordinate concepts	<i>inddele i, opdele i (subdivide into)</i>
co-ordinate concepts	<i>skelne mellem (distinguish between)</i> <i>adskille sig fra (differ from)</i>
comprehensive concept + partitive concepts	<i>bestå af (consist of)</i> <i>sammensat af (composed of)</i>
delimiting characteristics	<i>karakterisere ved, kendetegne ved (characterize by)</i>
intensional definitions	<i>definere som (define as)</i> <i>forstå ved (understand by)</i>

1. Cykler kan **inddeles i** hverdagscykler, sportscykler, transportcykler, HPV-cykler / liggecykler, børnecykler og en lang række andre typer.
(*Bicycles can be **subdivided into** everyday bicycles, sports bicycles, transport bicycles, HPV bicycles, children's bicycles, and a wide range of other types*)

‘Weak knowledge patterns’, on the contrary, are patterns that result in a high degree of noise, or patterns that result in findings with different types of terminological information requiring a lot of manual work on the part of the terminologist. An important example of the latter is the Danish verb *kalde* (*call*). Searching on this verb, one may identify terminological information such as terms, synonyms, relations among superordinate and subordinate concepts, and definitions or explanations. This has inspired me to investigate whether for the verb *kalde* (*call*), recurrent patterns exist over and above its valency pattern proper, i.e. patterns that might support a more precise identification of terminological categories. The study showed that hedges such as *også* (*also*), *ofte* (*often*), *almindeligvis* (*usually*), *tidligere* (*earlier*), *i dag* (*today*), and *undertiden* (*sometimes*) often co-occur with *kalde* (*call*). They turned out to be useful in validating the status of a particular term, i.e. whether it should be a synonym, a preferred term, or an obsolete term. In this way, a knowledge pattern such as *kalde* (*call*) combined with hedges also becomes a strong knowledge pattern for the relation between terms.

Applying knowledge patterns for terminology investigations, my earlier tests showed that a subdivision of the inventory into strong and weak knowledge patterns was advisable and also that the best search strategy was to begin by searching on strong patterns, which made it possible to predict which information categories would be the result [9,10]. For the testing of web-crawled corpora, the same strategy can be recommended. Similar results have been reached independently by Caroline Barrière [2].

Thus, the strong knowledge patterns of verbs have been applied for testing whe-

ther a web-crawled corpus has sufficient conceptual coverage. The approach may be criticized because individual knowledge patterns, not specific concepts, are the point of departure. Thus, no overall view of the knowledge patterns occurring together with a given concept will be obtained and therefore no full picture of the amount of conceptual information of a specific concept in a given corpus will be obtained.

6.2 Word Sketch Used for Identifying Knowledge Patterns

Word sketch for languages provided with high-level resources allows the terminologist to gain such an overview of the number of knowledge patterns and thus of the amount of potential conceptual information that may be related to a specific concept in the corpus. As shown in Fig. 3, word sketch contains information on verbs and related prepositions. In the case of *bicycle*, we find the strong knowledge patterns *classify as* and *class as* labeled ‘pp_obj_as’ in the utmost right-hand column. Besides, in the first column ‘object_of’, the verbs *define* and *classify* occur without a preposition.

7 Conclusion

The study aimed at finding a method for evaluating whether web-crawled corpora could be used for exam assignments in terminology. Focus was on linguistic and conceptual coverage as important criteria.

For languages provided with high-level resources in Sketch Engine, the degree of linguistic coverage can be tested by means of the term extraction function. The word sketch is another important support function allowing the terminologist to obtain an overview of the multi-word term candidates related to a specific term.

The degree of conceptual coverage has been tested by searching the corpus for strong knowledge patterns. The word sketch function has also proved well-suited for terminologists because it provides a quick overview of the knowledge patterns occurring together with specific concepts in a given corpus. Consequently, the functions of Sketch Engine can be usefully combined with knowledge patterns for evaluating web-crawled corpora on a qualitative basis, to make sure that the corpora comprise relevant terminological information in knowledge-rich contexts.

For languages not provided with high-level resources, however, the work process must be based on pragmatic, character string approaches. Searches for generic terms as part of composite terms often contribute to creating a good overview of the degree of linguistic coverage, especially when combined with the reduced frequency list, as shown in Fig. 2. To test the corpus for potential multi-word terms, concordance lists are used. To assess the degree of conceptual coverage, tests using searches for the knowledge patterns as the point of departure have been carried out, followed by manual linking of the concordances to the specific concepts. This approach shows that

right from the start, testing a Danish terminological corpus consisting of raw text only and carrying out subsequent terminology work will be a much more labor-intensive manual task than a similar task in a language provided with the high-level resources.

Finally, it turned out that many strong knowledge patterns for Danish identify semantic relations among concepts. Experience also shows that knowledge patterns often occur close to each other, and that the texts chunks in which knowledge patterns occur are often heavily loaded with conceptual information. If corpora contain these types of patterns, preliminary concept systems can be worked out on the basis of them. These are useful points of departure for exam assignments, since the first phases of the terminological process consist in analyzing relations among concepts and working out concept systems which in turn form the basis for drafting good definitions.

References

1. Baisa, Vít. "Sketch Engine for Terminology and Translation, webinar prepared for Term-Net." Accessed January 29, 2016.
<https://www.sketchengine.co.uk/category/news/>
2. Barrière, Caroline. "Semi-automatic corpus construction from informative texts." In *Lexicography, Terminology and Translation, Text-Based studies in honour of Ingrid Meyer*, edited by Lynne Bowker, 81-92. Ottawa: University of Ottawa Press, 2006.
3. Bowker, Lynne, and Jennifer Pearson. *Working with Specialized Language, A practical guide to using corpora*. London/New York: Routledge, 2002.
4. Daugaard, Jan, and Sabine Kirchmeier-Andersen. "The Odense Valency Dictionary Programme for Verb Coding." In *Odense Working Papers in Language and Communication No. 8*, edited by Jan Daugaard, 3-35. Odense: Odense University, 1995.
5. Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. "The Sketch Engine: ten years on." *Lexicography: Journal of ASIALEX*, volume 1 (2014): 7-36.
6. Kilgarriff, Adam. "Getting to know your corpus." In *Proceedings of The 15th International Conference on Text, Speech and Dialogue (TSD), Czech Republic*, edited by Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, 3-15. Berlin: Springer, 2012.
7. Meyer, Ingrid. "Extracting knowledge-rich contexts for terminography, A conceptual and methodological framework." In *Recent Advances in Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, 279-301. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2001.
8. Sketch Engine. "Writing Sketch Grammars." Accessed February 16, 2016.
<https://www.sketchengine.co.uk/writing-sketch-grammars/>
9. Weilgaard Christensen, Lotte. "Hvordan ord sporer termer og andre terminologiske oplysninger." In *Nordterm 2005: ORD OG TERMER, NORDTERM 14*, edited by Ágústa Þorbergisdóttir, 83-93. Reykjavík, 2006.
10. Weilgaard Christensen, Lotte. "Valency Patterns of Danish Verbs as Terminological Knowledge Patterns." In *Workshop Computational and Computer-assisted Terminology, LREC 2004, IV International Conference On Language Resources and Evaluation*, edited by Rute Costa, Lotte Weilgaard, Raquel Silva, and Pierre Auger, 20-23. Lisbon, 2004.

Modeling the dynamics of domain specific terminology in diachronic corpora

Gerhard Heyer¹, Cathleen Kantner², Andreas Niekler¹, Max Overbeck², Gregor Wiedemann¹

¹Leipzig University, Natural Language Processing Group, Computer Science Department

heyer|aniekler|wiedemann@informatik.uni-leipzig.de

²Stuttgart University, Chair for International Relations, Social Sciences Institute
cathleen.kantner|maximilian.overbeck@sowi.uni-stuttgart.de

Abstract: In terminology work, natural language processing, and digital humanities, several studies address the analysis of variations in context and meaning of terms in order to detect semantic change and the evolution of terms. We distinguish three different approaches to describe contextual variations: methods based on the analysis of patterns and linguistic clues, methods exploring the latent semantic space of single words, and methods for the analysis of topic membership. The paper presents the notion of *context volatility* as a new measure for detecting semantic change and applies it to key term extraction in a political science case study. The measure quantifies the dynamics of a term’s contextual variation within a diachronic corpus to identify periods of time that are characterised by intense controversial debates or substantial semantic transformations.

Keywords: Terminology extraction, semantic change, diachronic corpora, political science

1 Introduction

While the classical theory of terminology presupposes that key terms reflect objective, clear-cut concepts within static conceptual structures (Wüster 1979), recent advances in terminology work have highlighted the dynamics of terms in diachronic text corpora and propose explanations for the change and development of terms (S. Fernández-Silva et. al. 2011, Picton 2011). The methods for key term extraction in computational linguistics and terminology engineering can roughly be divided into *frequentist* and *Bayesian* approaches. On the one hand,

focusing on the frequency of terms, statistical tests such as log-likelihood-ratio can be employed to compare expected with observed term frequencies using reference corpora (Archer 2008). To detect changes in a term's usage, it is also common to observe a term's context and evaluate how it may change over time (Lenci 2008). By this approach, contextual variations can be measured using a bag of words document model and thresholds based on a tf/idf comparison of text stream segments (e.g. Kumaran and Allan, 2004). On the other hand, assuming a *Bayesian model* of topic and term distribution in documents, one can also use co-occurrence patterns and their local distribution in time to detect changing topics over time (Wang & McCallum 2006).

In most diachronic corpora, however, the patterns for the emergence of new terms, or contextual changes of existing terms, cannot be described just by reference to frequency or topic clusters (S. Fernández-Silva et. al. 2011). Rather, they are the result of a number of factors such as *centrality*, i.e. the use of terms and concepts to convey a change in the domain where the terms “all belong to a common topic in the domain and indicate an evolution in this topic” (Picton 2011, p. 147). Often, the increase or decrease of occurrences of terms in a domain is not related to novelty, but to the centrality/disappearance of a topic in the domain of application because of scientific or public discussion (ibid.).

In order to better describe and track controversial discussions reflected in diachronic corpora, we would like to introduce the notion of *context volatility*. Assuming a distributional model of meaning (Turney & Pantel 2010), we consider a term's global context (see below) as a second dimension for analyzing its salience and temporal extension in addition to term frequency. Changes over time in the global context of a term thus indicate a change of usage. Our novel approach differs from previous ones in the spirit of distributional semantics in important aspects: for us the *rate of change* is indicative of how much the “opinion stakeholders” agree, or disagree, on the meaning of a term. Fixing the usage of a term within a community of speakers seems in some ways similar to fixing the price of a stock at a stock market. Reversely, the analysis of the volatility of a term's global context can be employed to detect controversial or changing topics. In the following, we will first review related work on contextual variation of terms, and then explain the basic notions and assumptions of our approach. Finally, we will present first experimental results from a case study carried out in political science.

2 Context Change of Terms – Related Work

In terminology work, natural language processing, and digital humanities, several studies address the analysis of variation in context of terms in order to detect semantic change and the evolution of terms. Three different approaches to describe contextual variations can be distinguished: (1) methods based on the analysis of patterns and linguistic clues to explain term variations, (2) methods that explore the latent semantic space of single words, and (3) methods for the analysis of topic membership.

(1) Most studies in the area of terminology focus on particular terms, and look for linguistic clues and different patterns of variation in their usage to better understand the dynamics of terms such as Fernández-Silva, Freixa, and Cabré (2011) or Picton (2011). These studies take a particular term as starting point and inspect its neighbouring context to classify, analyse and predict changes of usage. In contrast, our approach takes a whole corpus as starting point, and aims at detecting terms that exhibit a high rate of contextual variation for some time.

(2) In NLP and digital humanities, distributional properties of text have been used to study the dynamics of terms in diachronic texts. Jatowt and Duh (2014) use latent semantics of words in order to create representations of a term's evolution. Hilpert (2011) proposed a similar method, which uses multidimensional scaling to find latent semantic structures, and compare them for different periods. These approaches try to model semantic change over time by setting a certain time period as reference point and comparing the latent semantic space to that reference over time. Terms can thus be compared with respect to their semantic distance or similarity over time. Again, our approach differs from these because we do not start with a fixed set of terms to study and trace their evolution, but rather we want to detect terms in a collection of documents that may be indicative of semantic change.

(3) Assuming a Bayesian approach, topic modeling is another method to analyse the usage of terms and their embeddedness within topics over time (Rohrdantz et al. 2011; Rohrdantz et al. 2012). These studies identify terms, which have changed in usage and context, and show that this change can be quantified by the probability of a term's membership in a topic cluster within the topic model used. Approaches like the one of Blei and Lafferty (2006) model the dynamics of a term's topic membership directly and allow the model to slightly change its co-

occurrence structure over time. Zang et al. (2010) modify hierarchical Dirichlet processes to measure the changing share of salient topics over time, and thus help to identify topics and terms that for are very prominent for some time. Jähnichen (2015) has extended this approach to identify topics that for some period of time contain rapidly changing terms, and thus can be considered to be indicative of conceptual changes. However, topic model based approaches always require an interpretation of the topics and their context. In effect, the analysis of a term’s change is always relative to the interpretation of the global topic cluster, and strongly depends on it. Topic models only generate a *macro view* on document collections. In order to identify contextual variations, we also need to look at the key terms that drive the changes at the *micro level*. Often these hot-button words fan the flames of a debate.

In sum, while related work on the dynamics of terms usually starts with a reference (like pre-selected terms, some pre-defined latent semantics structures, or given topic structures), we aim at automatically identifying terms that exhibit a high degree of contextual variation in a diachronic corpus. The typological category of *centrality* as introduced by Picton (2011) tries to capture the observation that central terms simultaneously appear or disappear in a corpus when the key assumptions, or consensus, amongst the stakeholders of a domain change. The measure of *context volatility* is intended to support exploratory search for such central terms in diachronic corpora, in particular, if we want to identify periods of time that are characterised by substantial semantic transformation. However, we do not claim that our measure quantifies meaning change or semantic change, the measure quantifies the dynamics of a term’s contextual information within a diachronic corpus.

3 Context Volatility - Intuition

Our focus for identifying context changes is on the retrieval of what *authors* consider “worth writing about” (for whatever reason). Any topic “worth writing about” represents some author’s point of view (at some point of time). On some topics there may be agreement, others may be contested – and this can change over time. “Hot-button” topics are highly controversial topics with a clear-cut distinction between proponents and opponents.

We observed that for competing opinion stakeholders, the linguistic context of key terms is *different*. For example, with the exception of the controversial term “nuclear power” and some stop-words, there is no overlap between the controversial positions on nuclear power based on excerpts from internet fora summarized below (table 1).

<i>Pro nuclear power</i>	<i>Contra nuclear power</i>
Nuclear power is a very efficient source of energy. It is also abundant, unlike fossil fuels (coal and oil).	Nuclear power plants are hard to control. Like in Fukushima 2011, a steam buildup in a nuclear reactor in Chornobyl, Ukraine, caused an explosion that released tons of radiation into contact with people and animals. The radiation released from nuclear fission is harmful to living organisms.

Table 1: Controversial positions on “nuclear power”

When dealing with real-life time-stamped data spanning long periods of time (e.g. newspaper texts, patent applications, or scientific publications), we observed, moreover, that the *global context* of terms does not need to be static, but may radically change. The global context of a term – we assume – consists of all its statistically significant co-occurrences within a corpus, where we measure significance using the log-likelihood ratio (Heyer et al., 2008).¹ To give an example, consider the changes in the global context of the German term “*Kredit*” (credit/loan) in the digital edition of the German weekly newspaper DIE ZEIT. Co-occurrence statistics computed on a yearly basis and visualized as context-networks display almost complete changes of the semantic context (see figure 1 for the graphs for the years 2005, 2007 and 2009).

¹ A term’s set of co-occurrences is computed on the basis of the term’s joint appearance with its co-occurring terms within a predefined text window taking an appropriate measure for statistically significant co-occurrence. The global context can also be displayed as a graph which contains the term and its context terms as nodes where the edges have a weight according to the significance value of the joint appearance of the terms.

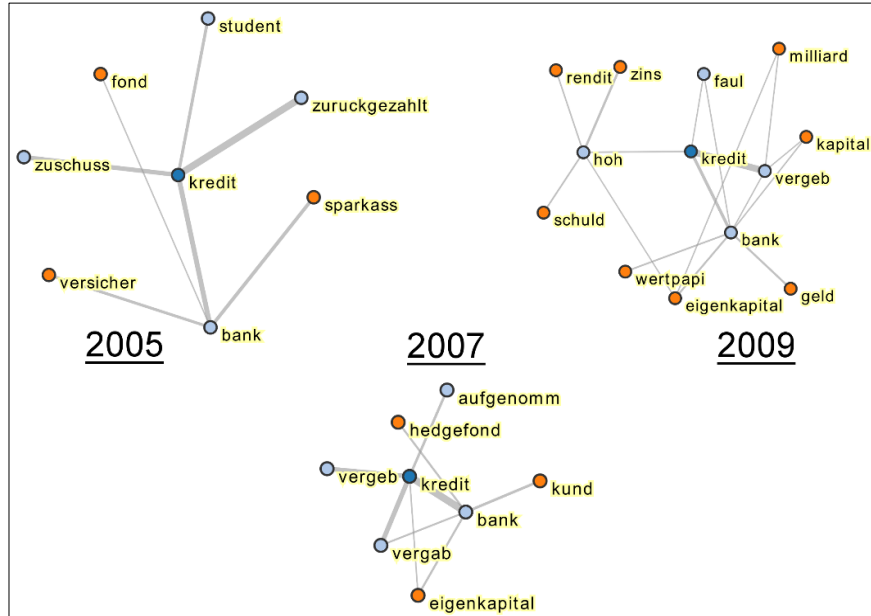


Figure 1: Changes in the global context of the German term “Kredit” (credit/loan)

While in 2005 the main usage apparently covered references in the context of student loans, in 2007 there is already a mention of net assets (*Eigenkapital*) in connection with credits granted by banks. Finally, in 2009, the modifier *hoch* (*high*) is linked to *Zins* (interest rates), *Rendite* (income return) and *Schuld* (debt). Furthermore, we see the evaluator *faul* (foul) linked to the word *Kredit* (loan). Quite obviously, the risks taken by banks granting bad credits was something worth reporting on, and by doing so, the global context of “*Kredit*” has changed substantially so that the link between *faul* and *Kredit* became almost collocational. Following this approach, a new multi-term expression can be viewed as a new term referring to the way banks were handling credits in 2009.

4 Context Volatility – Definition

The basis of our analysis is a set of time stamped text corpora, e.g. all editions of a digital weekly newspaper between January 2005 and December 2010 which is our test case in this paper. Our measure of the contextual changes is the mean *volatility* in the co-occurrence ranks of a

term. It is inspired by the widely used risk measure in econometrics and finance², and based on the ranking of significant co-occurrences in a defined time slice. A time slice is a set of documents belonging to a consecutive time span. The corpus is divided into time spans allowing, however, for various options from years, months, weeks, days or even hours to minutes. The example in this paper was created using months as the time spans of choice. Informally, we compute a term's change of context by averaging the changes in the *ranks* of its co-occurrences for a defined number of time slices. This can be conducted in a variety of ways. We considered all time slices in order to define a global measure of the dynamics of a term's context, e.g. the changes of its distributional semantics. Moreover, we also build the measure for a window of time slices for each term to produce a time series of a term's context change. *Context volatility* is then computed as the *average* of all rank changes of a term's co-occurrences for some period of time as follows:

1. Compute for every word w of the vocabulary V and every time slice t (days, weeks, years) in the data of all time slices T the set of co-occurrences, e.g. a term-term matrix C_t with co-occurrence weights for every time slice. The matrix has the dimension $V \times V$.³
2. Compute for every word the rank for every concurrent word for every time slice as a matrix $R_{V,T}$ where the rows represent the ranks of all co-occurrent words of w throughout the time slices. This matrix has the dimension $V \times T$ and is produced for every word in V .
3. Compute the *context volatility* of a word for a given history h in the time slices T by computing the difference between the 3rd and the 1st quartile of all ranks that the co-occurrences of word w take for all time slices in h , e.g. the interquartile range (IQR) of a row in $R_{w,T}$ where we limit the row to t elements of h . The result is again a matrix $CV_{w,T}$ where each row contains the IQR at a time slice t for a given history h .

² Yet, it is calculated differently and not based on widely used gain/loss measures. For an overview of miscellaneous approaches to volatility see Taylor (2007).

³ The weights can be set by significance measures like Log-likelihood, Dice, Mutual Information or a significance test based on the Poisson distribution. For this paper we used the log-likelihood significance measure.

4. Compute the global *context volatility* for a word w by averaging the columns, e.g. all co-occurents in $CV_{w,T}$, to compute the mean of all standard deviations in the rank changes. The result is a vector S_w which represents the quantity of context change as defined by the *context volatility* w.r.t the defined sequence of back-looking windows. If we define the back-looking history as the set of all time slices within the data, we get a single constant. If h is a window shorter than T we get a time series of quantified context changes for that term with the length $T-h$. In summary, we can define the final calculation of the volatility for h or T as

$$CV_{w,T} = \frac{1}{C_{w,T}} \sum_i IQR(Rank(C_{w,i}, T))$$

Here $C_{w,T}$ is the number of all co-occurences of w in T . $C_{w,i}$ is the i th co-occurence of w in T . $Rank$ represents a set of all ranks $C_{w,i}$ holds within T , and IQR is the interquartile range of those ranks.

As this computation is complex (at least $O(n^2 * t)$ with n the size of the vocabulary and t the number of time slices), we improved the runtime of our algorithm by considering only the overall most significant co-occurences (filtering out stop-words and pruning words with a document frequency < 3). We also used parallel computations to speed up the process. We parallelized the computation of the matrices C_t since they are totally independent from each other. Furthermore, we parallelized the computation of $R_{w,T}$, $CV_{w,T}$ and S_w to compute their values for every term separately. This way the whole process is scalable w.r.t T and V .

5 Use case – Issue Analysis in Political Science

The measure of context-volatility enables us to explore large amounts of documents and to identify periods of substantial semantic change. This opens fruitful ways for the identification of so-called “issues” in public political communication. A political issue is “a controversial social problem, which constitutes a broader topical structure, encompassing

several events as belonging together” (Kantner 2015, p. 40). Social problems are real-world matters involving a certain vocabulary. However, events, actors, opinions, cultural and technical features change over time. This results in a dilemma, especially when we want to identify issues over longer historical periods: On the one hand, we want to identify terms that characterize issues as some kind of generic social problem that at some points in time provoke intense and controversial discussions, and for which at different times different solutions have been proposed. On the other hand, we also want to identify those periods of time where the issue is being fed by new conflicts, contested, and redefined and thus undergoes semantic transformation. Therefore, we are interested in, both, terms that describe issues in general irrespective of contextual variation and semantic change, and at the same time exactly those terms that mark particular periods of crisis and semantic change within the issue.

In order to deal with that dilemma, we proceeded in two steps combining topic modeling with context-volatility analysis. Our use case is based on 397,729 articles from altogether 3,841 editions of the German weekly newspaper DIE ZEIT covering the period from 1946 – 2011.⁴ One central problem with standard topic-models is that they generate topics that are not intuitive and that they involve largely named entities such as people, places, and events. To compute the generic political issues for this document collection we, therefore, computed in a first step 30 topics based on the Latent Dirichlet Allocation Model (LDA) (Blei et al. 2003) after deleting all named entities such as names of people, places, and events. Since issues are defined as broader social problems, named entities referring to those people, places, and events, characterize an issue during a short time span. To delete the event-bias and to catch only the properties of the issues in general, the topic model was created without named entities.

Thirty topical fields could be distinguished. Among them, one topic relates to “*financial and economic policies*” (fig. 2). For the remainder of this paper we will focus on this relation as name for the topic represented by the 30 most probable terms inferred by the LDA model.

⁴ The data were retrieved from DTA corpus. The preprocessing of the text sources includes the following steps: sentence segmentation, tokenization, named entity recognition, multi-word unit identification, stop-word deletion, lower case transformation and lemmatization. The resulting term vectors for each sentence were used to create a sentence-term matrix for annual time slices. Those matrices were pruned to delete high frequent and low frequent words from the process. We used relative pruning and excluded vocabulary which is found in more than 99% and in less than 1% of the documents.

dollar, milliarde, jahr, prozent, geld, million, gewinn, zins, kredit, markt, fond,
 pfund, geschäft, kasse, bank, unternehmen, verlust, währung, investor, kunde,
 umsatz, anteil, konzern, schuld, investition, gold, verkauf, monat, versicherung,
 kauf

Figure 2: Words representing the topic “financial and economic policies”

When looking at the temporal salience of *financial and economic policies*, we clearly see changing phases of activity, e.g. high peaks in the early seventies relating to discussions of currency parities, or in 2008/2009 relating to the last financial crisis (fig. 3). The longitudinal data was produced by counting those documents within the corpus, which contain the context, e.g. topic (see fig. 2), from our inferred LDA model at a minimum of 30%.

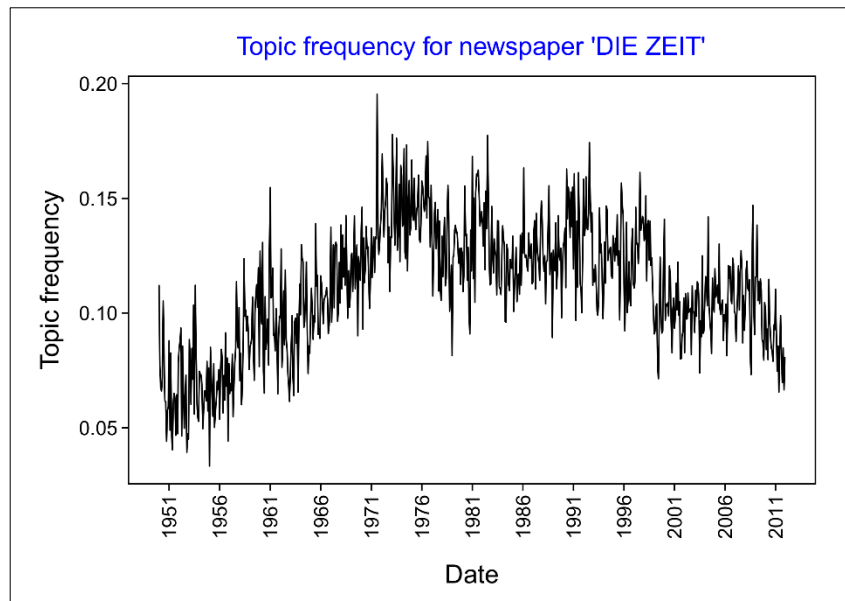


Figure 3: Temporal salience of topic, normalized, monthly basis

In order to identify issues in the technical sense, we then identified key terms within that topic that not only have a high relative frequency, or *tf/idf*-value⁵, but can also be considered to fuel controversial discussion. Thus, by looking for key terms in controversies, and by assuming that the context of these terms is rapidly changing, the measure of *context volatility* is a natural choice. In our case study, we wanted to test whether our *context volatility* measure is able to recognize the last financial crisis in 2008/2009. In order to do so, we applied the measure on a suitable sub-corpus of the whole data for one topic (*financial and economic policies*) and the years 2005 to 2010. This time we included the named entities again and, of course, we were pointed to some of these named entities that are characteristic for that period of time, and that describe the key actors of the crisis such as *Lehman Brothers*, or *Goldman-Sachs*. However, we also found terms like *Kredit* (loan), *Banken* (banks), *Fonds* (fonds) and *Schulden* (debts), that are constitutive of the general topic (fig. 4). Again, we constructed the global terms from the LDA model with the 30 most probable words from the topic. The top volatile terms created by our measure applied for all time slices (in months) of our corpus.

Important terms in financial and economic policies topic (1946-2011)	Top volatile terms in financial markets sub-corpus (2005-2010)
dollar, milliarde, jahr, prozent, geld, million, gewinn, zins, kredit, markt, fond, pfund, geschäft, kasse, bank, unternehmen, verlust, währung, investor, kunde, umsatz, anteil, konzern, schuld, investition, gold, verkauf, monat, versicherung, kauf	dollar, bank, kredite, anleger, unternehmen, geld, banken, schulden, wert, gewinn, umsatz, dresdner, lehman, goldman, zinsen, investieren, merkel, morgan, pfund, währungsfonds, wunder, zentralbank, aktien, estate, fonds

Figure 4: Comparison of global terms (topic) and top volatile-terms (context volatility over all time slices) in financial markets sub-corpus

From a methodological point of view, it is interesting to notice that *context volatility* of these terms highly correlates in time with intense public controversy, but not with the terms relative frequency. In figure 5, the *context volatility* and the relative frequency have been plotted for the

⁵ *Tf/idf* (term frequency / inverse document frequency) values represent the uniqueness and importance of terms within a document (Manning et. al. 2008).

terms *Kredit* and *Fond*. Both terms are good examples due to their strong context fluctuation within our exemplary issue. The ranges of values were aligned in order to overlay both longitudinal plots. We set a history h for the calculation of the *context volatility* of 6 months. The co-occurrence statistics were calculated for each month, which corresponds to monthly time slices. This means that we calculated a *context volatility* for each word at a time t based on the contextual changes from the last 6 month. The figures show that the relative word frequency does not correlate with the *context volatility*. Apparently, the possible change of context, the discursivity, salience, or centrality of a term, cannot fully be reflected by its frequency of usage. Longitudinal *context volatility* signals for terms, which in turn can be used to identify points in time where a semantic, or paradigmatic, change of the meaning of a term might happen. Further interpretations could be that the striking term is discussed from different points of view and *context volatility* thus reflects controversial discussion, or it can even be considered a weak signal for new adjustments within mainstream or established contexts. Of course, we can also calculate the volatility for the whole time span of the corpus highlighting terms, which appear in different contexts more often than other terms (see fig. 4).

For social scientists, the use of this measure of *context volatility* is highly profitable. With the growing accessibility of very large, long-time textual corpora, scholars are increasingly interested in (and dependent on) the use of automated textual analysis techniques in order to conduct comparative media studies, or to analyze parliamentary debates or presidential speeches. They want to grasp the salience of specific issues over time and among different countries. They are interested in identifying dominant discourses and frames of interpretation in public debates on issues such as immigration or foreign policy. Last but not least, they want to know which actors or organizations are the ones that are most visible in the media in light of important events such as the current refugee crisis in Europe or during the war in Libya 2011. In this regard, measuring the *context volatility* of terms or topics has pioneering character. Social scientists so far could only come to terms with these questions by measuring the frequency of key terms, term or collocation lists, or topic models. However, by measuring the volatility of co-occurring word contexts, they can now approach a second crucial dimension to determine the salience of an issue: The degree of *contentiousness* of a specific term or topic. Assuming that an issue can

be understood as an ongoing flow of communication on matters, which are controversially discussed among different stakeholders, it can be concluded that a topic is not only relevant because it is highly frequent in a given amount of textual data. Hot-button issues might moreover be characterized by high variance of their linguistic contexts. Public stakeholders, due to their different views on the same subject, use different terminologies and try to push their opinion in the public contest of opposing convictions. Thus, as depicted in figure 5, it is important to consider both – frequency as well as *context volatility* – in order to best determine the salience of an issue or term. Otherwise, the importance of those terms that are highly frequent might be overrated while the salience of those (even low frequent) terms that have a high degree of *context volatility* are neglected.

6 Conclusion

In this paper, we introduced the notion of *context volatility* as a new measure to identify semantic change of key terms and issues in specialized domains of discourse. Our case study in the field of political science focusing on the analysis of political issues demonstrated the usefulness of this measure. It was possible to identify controversial issues marked by certain key-terms that are in general characteristic for the issue as well as some key-terms highly dependent on particular circumstances and crisis situations– such as *Kredit* or *Fond* for the last financial crisis 2008/09. Yet, the usefulness of the new measure of *context volatility* is, of course, not restricted to this area of application. Because it helps to distinguish clearly between the frequency and contextual usage of terms, it may also be of use in other domains of scientific analysis, the identification of new terms in marketing studies, or technology mining, and terminology extraction in diachronic corpora – especially in cases where rather static standard methods prove to be unable to deal with semantic change.

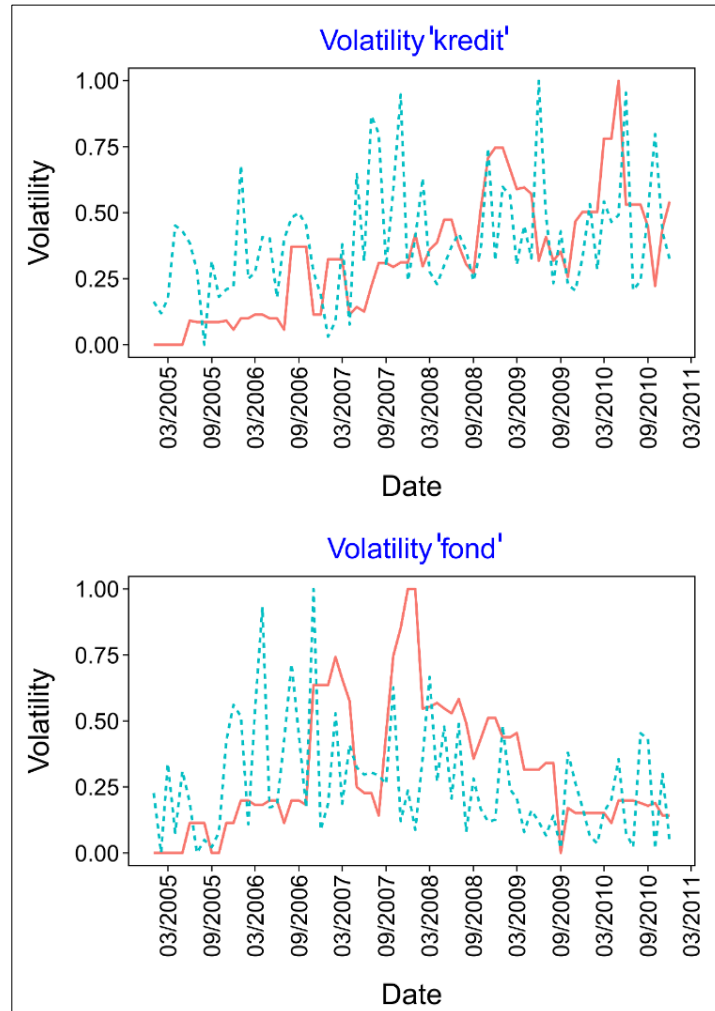


Figure 5: Relative term frequency (dotted line) and volatility of key terms (solid line)

References

Archer, D., (ed.)(2008). What's in a word-list? Investigating word frequency and keyword extraction. Ashgate, Aldershot.

Blei, D. M., Ng, A.Y., and Jordan, M.I. (2003). *Latent dirichlet allocation*. The Journal of Machine Learning Research 3: 993–1022.

Blei, D. M. and Lafferty, J. D. (2006). *Dynamic topic models*. In Proceedings of the 23rd international conference on Machine learning, 113–120.

Harris, Z. (1954). *Distributional Structure*, Word 10 (23): 146-162.

Heyer, G., Quasthoff, U. and Wittig, T. (2008): *Text Mining – Wissensrohstoff Text: Grundlagen, Algorithmen, Beispiele*, Bochum, w3l-Verlag.

Hilpert, M. (2011): *Dynamic Visualizations of Language Change: Motion Charts on the Basis of Bivariate and Multivariate Data from Diachronic Corpora*. International Journal of Corpus Linguistics 16 (4): 435–61.

Jatowt, A. and Duh, K. (2014): *A framework for analyzing semantic change of words across time*. In Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries, 229–38.

Jähnichen, P. (2015): *Topics over time – A new approach to dynamic topic models*, Ph.D. Thesis, Leipzig University.

Kantner, C. (2015): *War and Intervention in the Transnational Public Sphere: Problem-Solving and European Identity-Formation*. London, Routledge.

Kumaran, G.; Allan, J. (2004): *Text classification and named entities for new event detection*. In SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pages 297–304, New York, NY, USA. ACM.

Lenci, A. (ed.) (2008): *From context to meaning: distributional models of the lexicon in linguistics and cognitive science*, Italian Journal of Linguistics, 20(1): 1-31.

Manning, C. D., Raghavan P. and Schütze H. (2008): *Introduction to Information Retrieval*. New York: Cambridge University Press.

Picton, A. (2011): *Picturing Short-Term Diachronic Phenomena in Specialised Corpora. A Textual Terminology Description of the Dynamics of Knowledge in Space Technologies*. Terminology, 17(1), 134-156.

Fernández-Silva, S., Freixa J. and Cabré, M. T. (2011): *A proposed method for analysing the dynamics of cognition through term variation*. Terminology 17(1). p. 49-73. Amsterdam: John Benjamins.

Taylor, S. (2007): *Asset Price Dynamics, Volatility, and Prediction*. Princeton and Oxford, Princeton University Press.

Turney, P. D. and Pantel, P. (2010): *From Frequency to Meaning*. Vector Space Models for Semantics, in: Journal of Artificial Intelligence Research 37: 141-188.

Rohrdantz, C., Hautli A., Thomas Mayer, Miriam Butt, Daniel A. Keim, und Frans Plank (2011): *Towards Tracking Semantic Change by Visual Analytics*. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, 305–310.

Rohrdantz, C., Niekler, A., Hautli A., Butt M. and Keim, D. A. (2012): *Lexical Semantics and Distribution of Suffixes: A Visual Analysis*. In Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, 7–15.

Wang, X. and McCallum, A. (2006): *Topics over time: a non-Markov continuous-time model of topical trends*. In KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 424–433.

Wüster, E. (1979): *Einführung in die Allgemeine Terminologielehre und Terminologische Lexikographie*. Viena, Springer.

Zhang, J. et. al. (2010): *Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora*. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.

Self-tuning ongoing terminology extraction retrained on terminology validation decisions

Alfredo Maldonado and David Lewis

ADAPT Centre*, School of Computer Science and Statistics
Trinity College Dublin
Ireland

{alfredo.maldonado,dave.lewis}@adaptcentre.ie

Abstract Automatic terminology extraction (ATE) is a first step in many terminology management processes. When applied on content, a linguistic pattern filter and statistical ranker (a “filter-ranker”) produces a ranked list of term candidates to be manually reviewed and validated by a terminologist. Ongoing content creation demands the re-application of ATE on each batch of new content. Unfortunately, traditional filter-rankers cannot learn from previous terminology validation decisions. This paper shows that it is feasible to replace traditional filter-rankers with a machine-learning terminology extraction method that is able to “self-tune” based on terminologists’ validations and is thus suitable for ongoing ATE as new content is created. Not only does this method perform better than traditional filter-rankers, but by taking advantage of the manual validation process already in place in terminology extraction workflows, it is possible to improve on similar machine learning systems that are trained only once on a static corpus but are used repeatedly on new content.

Keywords: terminology extraction, ATE, SVM, ACL RD-TEC

1 Introduction

Automatic terminology extraction (ATE) is a first step in the terminology processes associated with many content creation, curation and translation projects. An ATE system is expected to extract a list of specialised, technical or corporate *key* terms from a given corpus or document collection. Terminologists then research, validate, define, manage and/or translate these extracted terms, depending on the actual goal of the ATE effort. The users of the validated and curated glossary (or terminology) are authors, translators, marketing professionals, and other content-creation workers who seek to adhere to organisational, corporate, professional or institutional terminology standards. Style checking and language quality software as well as machine translation systems can also consume this terminology.

Terminology extraction, however, is rarely a once-off affair. As new content gets created, new products, services, features, processes, concepts and other pieces of knowledge get created, all potentially requiring new terminology. Failing to extract terminology at regular intervals along the content creation life-cycle runs the risk of missing the majority of terms by the time the content has grown significantly. This can be shown through the ACL RD-TEC¹ corpus version 1.0 (QasemiZadeh and Handschuh, 2014; see Sec. 4), a collection of Association of Computational Linguistics (ACL) academic papers written between 1965 and 2006, in which specialised terminology has been manually identified and annotated. Figure 1 shows

* The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

¹ http://atmykitchen.info/datasets/acl_rd_tec/. The ACL RD-TEC corpus itself was built upon the ACL ARC corpus (Bird et al., 2008) which was in turn derived from the ACL Anthology <http://aclanthology.info>.

the proportion of new terms identified in each year's worth of ACL papers and the trend of these proportions across the years. The black dots show the proportion of new terms at each year whilst the blue continuous line depicts the trend, with shaded areas indicating the trend's statistical confidence bounds. For the purposes of this graph, a term is counted as a new term in a given year, if it occurs in at least one paper written in that year and has not been featured in any paper written in any prior year². The proportion of new terms in a given year is the number of new terms in that year divided by the total number of terms in that year.

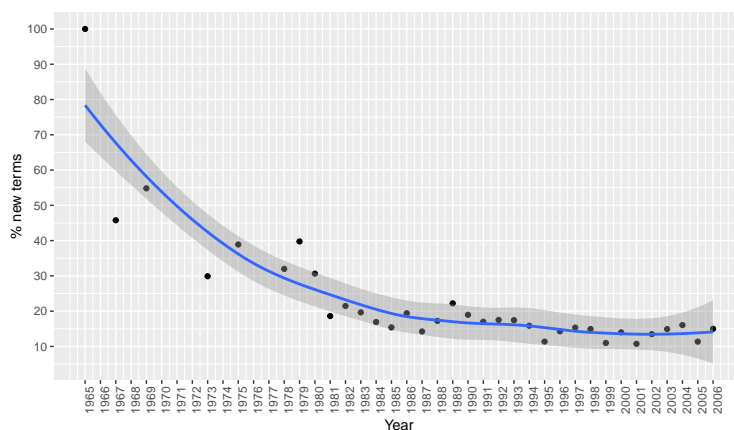


Figure 1. Proportion of new terms per year in the ACL RD-TEC corpus

Notice how the proportion of new terms drops dramatically within the first few years as one would expect. However, the proportion of terms never reaches zero in the long run. Notice that since 1981 the proportion of new terms every year remains relatively constant, fluctuating between 12% and 25%. This implies that failing to do terminology extraction in a subsequent new year will result in missing a substantial amount of new terms. Within a few years, the number of new terms missed will have exceeded the number of terms already captured in all previous years.

As a result, terminology extraction must be conducted repeatedly at appropriate intervals in order to optimally capture the terminology produced by a content-creating organisation, company or professional community. Unfortunately, the standard terminology extraction tools, usually a combination of linguistic pattern filters and statistical rankers (“filter-rankers”), do not readily lend themselves to ongoing use since they will output old or previously extracted terms mixed with new terms. One solution is to automatically filter out previously extracted and/or processed terminology in subsequent runs of a traditional filter-ranker, as proposed by Warburton (2013). However, the filter-ranker will still not be able to generalise from previous validation exercises in order to improve on the quality of newly extracted/ranked terms. In other words, there is no learning (or retraining) feedback loop across successive ATE iterations.

Since terminologists usually review (and validate) the output of terminology extractors, this paper proposes to use the reviewed/validated list of term candidates as feedback to a machine learning terminology extractor (MLTE). The manual effort dedicated during a terminology

² As 1965 is the first year in the corpus, 100% of the terms occurring in this year are considered new. There are some years missing a dot (1966, 1968, etc.) This is because no ACL conferences were held in those years.

extraction iteration can thus be used to improve the extractor’s performance in the next iteration. By performing simulations of terminology extraction-validation-retraining cycles, this paper shows (Sec. 5) that an MLTE trained in this manner outperforms MLTEs trained only once on a static corpus, as well as traditional filter-rankers.

This paper is structured as follows. Section 2 surveys previous work on ATE, especially on filter-rankers and methods based on machine-learning. It also describes the relationship between our work and Warburton’s (2013) proposal. Section 3 presents the general methodology for an MLTE system exploiting the terminology extraction-validation-retraining cycle whilst Section 4 describes how this methodology was adapted for the simulation experiments discussed in Section 5. These simulation experiments were conducted using a subset of the terminology-annotated ACL RD-TEC corpus. As the papers in this corpus are time-stamped, it is possible to group papers in chronological batches. In our experiments, terms are extracted from each chronological batch and compared with the valid terms occurring in the papers belonging to that batch as per the corpus’ own terminology annotation, thus simulating the manual validation process. Section 5 also offers conclusions and presents avenues for future research.

2 Previous work

The *de facto* standard method for ATE is a two-step approach. The first step, called the linguistic filtering step, involves identifying sub-sentential fragments from text that satisfy a syntactic pattern from a list curated by a terminologist, domain expert or computational linguist (Nakagawa, 2000; Pazienza et al., 2005). Typical patterns are noun+noun as in *terminology extraction*, adj+noun+noun as in *statistical machine translation*, among others. The second step, the statistical ranking step, ranks the candidates identified in the first step according to some statistical score acting as a proxy for their *termhood* and *unithood* in the text (Kageura and Umino, 1996). Statistical scores include mutual information (Church and Hanks, 1990), log-likelihood ratio of association (Dunning, 1993; McInnes, 2004), Pearson’s Chi-Squared test, Student’s t-score test, Fisher’s exact test (Pedersen, 1996; Purandare, 2004, pp. 35-48), TF-IDF (Spärck Jones, 1972), the C-Value/N-Value/NC-Value methods (Frantzi and Ananiadou, 1996; Frantzi et al., 2000) and even raw term frequency.

There have also been contrastive methods that compare the distribution of a term candidate in a domain-specific corpus against its distribution in a general-language corpus or a corpus from another domain. Methods include relative frequency ratios (Damerau, 1993; Ahmad et al., 1999), domain consensus, domain relevance (Navigli and Velardi, 2002), lexical cohesion of multi-word term candidates³ (Park et al., 2002; Sciano and Velardi, 2007), etc. In addition, researchers have created composite rankers based on linear combinations of several statistical features, reporting performance improvements over single-feature rankers (Loukachevitch, 2012; Bolshakova et al., 2013).

A third step, often not explored in detail in the ATE literature, consists in manually reviewing the ranked list of term candidates, usually concentrating on the top n ranking candidates or those scoring above some threshold. A subset of these term candidates deemed to be valid terms is then further processed in the terminology pipeline. For the purposes of this paper, this step shall be called the terminology validation step and is conducted by a terminologist, a person who is either a professional terminologist or a trained person with sufficient domain knowledge and linguistic experience to carry out this task competently. Not many works aim to take advantage of this manual validation step with the view of improving future term extraction-validation exercises, despite the development of numerous supervised machine learning approaches that depend on manually-annotated datasets, such as Pecina and

³ In this work, the lexical cohesion of multi-word term candidates is called *term cohesion*.

Schlesinger (2006), Pecina (2010) and QasemiZadeh et al. (2012), who built linear classifiers based on either logistic regression or support-vector machines (SVMs). A notable exception is Warburton (2013), who proposes to curate exclusion lists of several kinds such as a general lexicon list, a list of already known valid terms, a list of noisy items, etc. These lists are created by a terminologist from the terms automatically extracted using a filter-ranker during the validation of such terms. A computer program keeps track of the validation decisions for each candidate made by the terminologist and thus adds each candidate to the appropriate list. In a subsequent extraction-validation cycle, the terminologist can use these exclusion lists to automatically filter out term candidates produced by the filter-ranker, thus considerably reducing the amount of manual work associated with the validation step. The present paper seeks to combine Warburton’s approach with the machine learning classification approach. Instead of curating exclusion lists, we only ask the terminologist to label each extracted candidate as either a valid or a non-valid term via some user interface. The set of labelled term candidates is then used as training data for an SVM classifier.

3 Methodology

The self-tuning ongoing MLTE method sketched at the end of Section 2 can be described more formally as follows. A set of term candidate n-grams $C_t = \{c_1, \dots, c_n\}$ are extracted and counted from a given batch (set) of documents $b_t = \{d_1, \dots, d_m\}$ available at some point in time t . For each extracted n-gram $c_i \in C_t$, a record of its part-of-speech (POS) pattern as returned from some parser⁴ is also kept: $POS_t = \{pos(c_1), \dots, pos(c_n)\}$. Given a classifier $f(V_p)$ trained on a set of past validations $V_p = V_{t-k} \cup \dots \cup V_{t-1}$ from the last k recent batches b_{t-k} to b_{t-1} , each term candidate n-gram c_i is predicted to be a valid term or not by $y_i = f(V_p; c_i)$, where each y_i is a binary label indicating whether term candidate c_i is a valid term (1) or not (0). Those candidates selected (i.e. predicted to be valid) by the classifier are then presented to a terminologist who based on his/her expertise will either confirm or change the validity status of these candidates. This is the so-called validation step. The values confirmed and changed by the terminologist constitute the validations V_t for the current batch b_t , which will be used, along with other recent validations $V_n = V_{t-k+1} \cup \dots \cup V_t$ to train a new classifier $f(V_n)$ to be used to select term candidates from a batch b_{t+1} to be available at a future time $t + 1$. The collected POS patterns are used to filter out term candidates automatically⁵. Those c_i candidates that have a POS pattern $pos(c_i)$ not shared with at least one valid term from the recent batches b_{t-k} to b_{t-1} are automatically excluded before prediction.

Notice that this method does not require to curate term lists and does not require that terminologists craft any POS pattern filters *a priori*. Notice as well that the number of recent batches to consider, k , for training the classifier makes old data expire automatically. In addition, the method still allows the usage of supplementary filters, for example, excluding terms that are present in the current batch but that are already captured in the terminology database.

4 Experimental setup

4.1 Extraction-validation simulation

The terminology extraction-validation methodology presented in Section 3 is simulated by dividing a subset of the ACL RD-TEC corpus in separate chronological batches following

⁴ In this work we use the Stanford Parser (Chen and Manning, 2014) which is applied to full sentences where each candidate n-gram appears.

⁵ Whilst this filtering is optional, it is recommended as data points will be highly skewed towards the non-valid class ($y = 0$). This filtering reduces this skewness somewhat.

the corpus' own time-stamping encoded in the paper filenames. Such paper filenames include C04-1001_c1n.xml, J04-1003_c1n.xml and P04-1027_c1n.xml, where the first letter is a code for the ACL journal or conference (i.e. "the venue") where the paper was published and the following two digits indicate the year of publication (2004 in these cases). This is a convention established in the original ACL Anthology. The subset used in the simulation experiments are the 2,781 papers published from 2004 to 2006, gathering a total of 9,114,767 words with each paper tallying an average of 3,300 words. To better simulate extraction-validation iterations conducted at regular intervals, these papers are divided in chronological batches of comparable sizes. Each batch contains at most 40 papers from a single year and venue. On average each batch contains 36.6 papers. This yields a total of 69 separate batches.

The simulation is started by extracting all n-grams ($1 \leq n \leq 7$) from batch b_1 papers. These n-grams are then matched against the valid term⁶ annotation in ACL RD-TEC. Non-valid term n-grams that do not share a POS pattern with at least one valid term are automatically filtered out. This set of valid terms and retained non-valid terms constitute the training set for the SVM classifier (actual features described in Sec. 4.2). Then, n-grams are extracted from batch b_2 papers. Those b_2 n-grams that also occurred in b_1 are automatically filtered out. Those b_2 n-grams that do not share a POS pattern with b_1 valid terms are also removed. The remaining set of b_2 n-grams is the test set for the classifier. The classifier trained on b_1 data is then used to predict the validity of each term in the b_2 test set. Evaluation is done by comparing the annotation of each term in the test set with its predicted value (Sec. 4.4). This process is repeated for b_3 , except the training data used includes terms from b_1 and b_2 . In general, the training data for batch b_t will be the union of terms from batches b_{t-k} to b_{t-1} , where k is the history size, the number of past batches we want to consider as training data. It should be pointed out that regardless of the value of k , terms that occurred in all previous batches (from b_1 to b_{t-1}) are always excluded from b_t 's test set as we want to extract new terms only.

4.2 Features

Combinations of several of the statistical features presented in Sec. 2 were explored in preliminary experiments. Based on these, the features selected for the final experiments are two formal binary features, **POS pattern** and **character 3-gram**⁷, aimed at making the classifier sensitive to the typical syntactic patterns and morphological shapes of valid terms, as well as two sets of domain contrastive features aimed at detecting terms that are typical of the specialised domain of interest (computational linguistics in this case) and atypical of other (contrastive) domains. These domain contrastive features are:

- **Domain relevance (DR)** (Navigli and Velardi, 2002) measures the degree to which a term t is relevant to domain D_i . It is defined by (1) where $P(t|D_i)$ is estimated by (2).

$$DR(t, D_i) = \frac{P(t|D_i)}{\sum_{j=1}^n P(t|D_j)} \quad (1) \quad P(t|D_i) = \frac{f(t, D_i)}{\sum_{s \in D_i} f(s, D_i)} \quad (2)$$

- **Term cohesion (TC)** (Park et al., 2002; Sclano and Velardi, 2007) seeks to measure the degree of cohesion of multi-word term candidate t in a domain D_i :

$$LC(t, D_i) = \frac{l(t)f(t, D_i) \log f(t, D_i)}{\sum_{w_j \in t} f(w_j, D_i)} \quad (3)$$

where $l(t)$ is the length (number of words) of term candidate t , w_j are the individual words making up t and $f(x, D_i)$ is the frequency of word or term x in domain D_i .

⁶ ACL RD-TEC distinguishes between technology and non-technology valid terms. This distinction is not made here.

⁷ Notice these character 3-grams are only used as features for the classifier. The extracted term candidates are word n-grams of sizes $1 \leq n \leq 7$. This word n-gram size is a parameter set by the user.

In practice, D_i is some corpus belonging to a particular domain. In this work we model domains from two sources. One is a 2009 dump of Wikipedia⁸ clustered into 500 clusters using CLUTO (Karypis, 2003). Each cluster is interpreted to be a topical domain of Wikipedia from which DR and TC scores are computed for a term candidate, yielding two real-valued subvectors of 500 dimensions each. The other source is the history of batches up to the current batch (b_1 to b_t) from the same ACL RD-TEC sample, which is also clustered using CLUTO into c clusters, where c is 2.5% of the number of papers in the history. For each of these clusters DR and TC scores are also computed, yielding another set of real-valued subvectors of c dimensions each.

For each term candidate, the different feature subvectors computed (POS patterns, character 3-grams, Wikipedia and batch-history DR and TC) are concatenated to form a single vector, which is then L^2 -normalised.

4.3 Experiments

Three types of experiments are conducted: two baselines and our approach.

- **Baseline 1: Standard single-feature filter-rankers.** We extract term candidates from each batch b_i using the rankers implemented in the JATE Toolkit⁹ (Zhang et al., 2008), using n-gram extraction. This baseline follows the same protocol described in Sec. 4.1, except that instead of using a classifier, we use a single-feature statistical ranker. The filtering described in that simulation is also taking place in this baseline. So, essentially, this baseline simulates the process suggested by Warburton (2013) by keeping exclusion lists of previously extracted valid terms and terms with a POS pattern not associated with valid terms.
- **Baseline 2: SVM trained on first batch.** We train the SVM classifier on the first batch and use that classifier to extract terms for all subsequent batches. This baseline simulates the case of an extractor trained once and re-used in all subsequent batches with no retraining.
- **Our approach: SVM trained on last k batches.** We conduct the extraction-validation simulation as described in Sec. 4.1 while trying different history sizes k exhaustively.

The SVM experiments employ the LIBLINEAR SVM classifier (Fan et al., 2008).

4.4 Evaluation

Performance is assessed using **precision** and **recall**, which are the standard measures of classifier performance evaluation:

$$P = \frac{\text{valid terms selected}}{\text{total selected terms}} \times 100\% \quad (4) \qquad R = \frac{\text{valid terms selected}}{\text{total valid terms}} \times 100\% \quad (5)$$

A term is deemed to be selected (or extracted) if it is predicted to be a valid term by the classifier. Recall measures the coverage of our extractor (i.e. the percentage of valid terms in a batch that were selected). A low recall value indicates that the extractor is missing many terms. Precision on the other hand measures the percentage of true valid terms from the selected terms. A low precision value indicates that our extractor has produced many false positives. In ATE, we want to identify as many true valid terms as possible, potentially at the risk of having a relatively high number of false positives. So we are interested in achieving high recall at the expense of a moderate precision.

⁸ Wikipedia dump kindly cleaned, pre-processed and made available by the Web as a Corpus initiative (Baroni et al., 2009) at <http://wacky.sslmit.unibo.it/doku.php?id=corpora>

⁹ <https://code.google.com/archive/p/jatetoolkit/> and <https://github.com/ziqizhang/jate>

Filter-rankers, like those used in the Baseline 1 experiments, cannot be evaluated directly by precision and recall. However, by considering the top N ranked candidates as valid term predictions and all other candidates as non-valid term predictions, one effectively converts a ranker into a classifier that can be evaluated using standard precision and recall (Pecina, 2010). The problem then becomes finding an appropriate value N . If a batch test set contains v valid terms a perfect ranker will return all of these valid terms in the first v positions. So, by setting $N = v$ we will achieve a precision and a recall of 100%. In reality however, the ranker will be expected to be less than perfect. A solution could be to set $N = 2v$, giving the ranker twice the chance of finding all of the valid terms. Notice though that while a ranker will still be able to score a maximum of 100% recall under these conditions, it can only expect to obtain a maximum of 50% precision. Results for four rankers implemented by JATE (C-Value, GlossEx, Raw Term Frequency and Weirdness¹⁰) are presented in Figure 2. The plots on the left show Precision and Recall when using the conversion $N = 2v$ and the plots on the right show the same performance measures but for the $N = 7v$ conversion strategy. In all plots, the x axis represents the batch being evaluated. Trend lines are also included in the plots. Section 5 discusses these results.



Figure 2. Performance of filter-rankers implemented in JATE (Baseline 1) using $N = 2v$ (left) and $N = 7v$ (right)

5 Results and conclusions

The terminology extraction-validation simulation presented in this paper requires a parameter to be set manually: the history size k , i.e. the number of past batches (b_{t-k} to b_{t-1}) to use as training data for current batch b_t . We conducted systematic experiments trying out each possible history size ($1 \leq k \leq 68$) and found that our method performed quite similarly for most sizes, except for the smallest (1 and 2) which performed slightly worse. We found that on average the best performing history size was 16, which yielded an average recall of 74.17% across all batches. Accordingly, Figure 3 shows results using our method with a history size of 16 (ONGOING) along with the two baselines described in Sec. 4.3 (B1-CVALUE and B2-STATIC). For Baseline 1 only the C-Value ranker is plotted in Figure 3 as this is one of the most widely used rankers.

¹⁰ The Weirdness score (Ahmad et al., 1999) is a relative frequency ratio score closely related to equation (2).

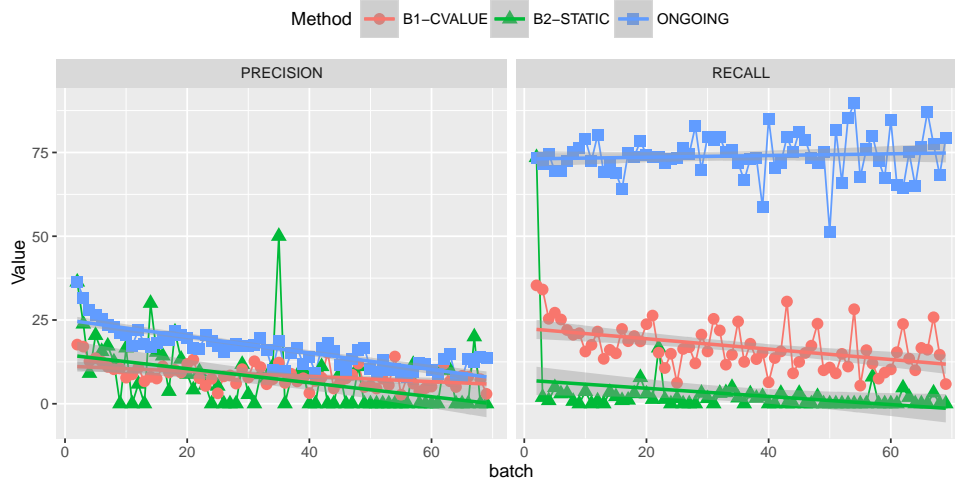


Figure 3. Performance of two baseline methods and our self-tuning ongoing method

The first thing to notice from this figure is that precision scores on the left-hand side tend to be far worse than the recall scores on the right-hand side. This means that all three methods produce a considerable amount of false positives, unfortunately. Our method (ONGOING), however produces the highest recall of all three methods by far, and while fluctuating, the recall trend is slightly on the increase. Compare this performance to that of B2-STATIC, the SVM trained on the first batch and used to extract terms in all subsequent batches with no retraining after each validation. Recall in B2-STATIC performs very well at the beginning (matching ONGOING), but it sinks to the bottom very quickly and largely remains there. This shows that the retraining conducted after each validation is indeed keeping the ONGOING classifier sensitive to new terms. If this retraining stops, the performance drops. So, a terminology extractor trained on a static set of data will not be able to perform well in the long run.

The recall of the C-Value ranker (B1-CVALUE) oscillates between 0 and 30%. Whilst performing better than B2-STATIC, its performance is far more modest than ONGOING, despite the usage of filtering/exclusion lists. In fact, this filtering is perhaps detrimental to the performance of all JATE's filter-rankers. Notice in Figure 2 that both precision and recall decrease without exception as we move towards newer batches. It is possible that the rankers tend to find terms in subsequent batches that are removed by the old term filtering mechanism employed, leaving few good candidates to report. Since the rankers have no information as to the terms being filtered out, they are unable to re-weight the remaining term candidates in order to maintain a high recall level. Whilst filter-rankers can be invaluable in extracting terminology in a new project, they are not suited for ongoing terminology extraction, even when automatic filtering of old terms is implemented, because they lack a feedback loop mechanism.

The results presented in this paper are based on a simulation. In a real terminology validation process, the terminologist has the discretion and flexibility to examine more or fewer term candidates depending on their experience, the quality of terms returned by a classifier or ranker, the size of the batch, among other factors. So, a more realistic, human-based study of a filter-ranker vs. a retrained MLTE is warranted. However, this work does demonstrate that the

manual validation process already in place (implicitly or explicitly) in virtually all terminology extraction tasks can be used effectively to retrain machine-learning methods to improve the quality of the extraction itself. In most if not all situations, an extractor based on such a machine-learning method could be readily used as a drop-in replacement for filter-rankers already in place in existing terminology workflows.

For future research we plan to conduct human-based benchmarks in order to confirm the simulations presented here. We also plan to address the low precision scores reported by exploring new features and post-processing strategies like re-ranking the candidates output by the classifier using traditional and new terminology ranking algorithms. Whilst the experiments here focused on one particular dataset from one particular domain (Computational Linguistics academic papers), the method should be applicable to other time-stamped datasets from other domains. So future research will also explore whether classifiers can rely on domain-independent features or whether they must depend on domain-specific features. Finally, the role of the contrastive corpus should be further investigated. Here we employed Wikipedia. However, many, very specific terms will not feature in Wikipedia at all. So we must find a way to cope with those cases. One way would be using a fallback strategy such as relying on sub-terms that do appear in the contrastive corpus. Another would involve using language modelling techniques or distributional MWE composition techniques in order to estimate the values of features of terms missing in the contrastive corpus.

References

- Ahmad, K., L. Gillam, and L. Tostevin (1999). University of Surrey participation in TREC8: Weirdness indexing for logical document extrapolation and retrieval (WILDER). In *Proceedings of the Eighth Text Retrieval Conference*.
- Baroni, M., S. Bernardini, A. Ferraresi, and E. Zanchetta (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3), 209–226.
- Bird, S., R. Dale, B. J. Dorr, B. Gibson, M. T. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. Radev, and Y. F. Tan (2008). The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakesh, pp. 1755–1759.
- Bolshakova, E., N. Loukachevitch, and M. Nokel (2013). Topic Models Can Improve Domain Term Extraction. *Advances in Information Retrieval. 35th European Conference on IR Research (ECIR 2013). Lecture Notes in Computer Science*. 7814, 684–687.
- Chen, D. and C. D. Manning (2014). A Fast and Accurate Dependency Parser using Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, pp. 740–750.
- Church, K. W. and P. Hanks (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1), 22–29.
- Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management* 29(4), 433–447.
- Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1), 61–74.
- Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin (2008). LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning* 9(2008), 1871–1874.
- Frantzi, K. T. and S. Ananiadou (1996). Extracting nested collocations. In *Proceedings of the 16th International Conference on Computational Linguistics -Volume 1*, Copenhagen, pp. 41–46.
- Frantzi, K. T., S. Ananiadou, and H. Mima (2000). Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries* 3(2), 115–130.
- Kageura, K. and B. Umino (1996). Methods of Automatic Term Recognition: A Review. *Terminology* 3(2).
- Karypis, G. (2003). CLUTO - a clustering toolkit. Technical report, Department of Computer Science, University of Minnesota, Minneapolis, MN.

- Loukachevitch, N. (2012). Automatic Term Recognition Needs Multiple Evidence. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, pp. 2401–2407.
- McInnes, B. T. (2004). *Extending the log likelihood measure to improve collocation identification*. Msc., University of Minnesota.
- Nakagawa, H. (2000). Automatic term recognition based on statistics of compound nouns. *Terminology* 6(2), 195–210.
- Navigli, R. and P. Velardi (2002). Semantic Interpretation of Terminological Strings. In *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE 2002)*, Nancy, pp. 95–100.
- Park, Y., R. J. Byrd, and B. K. Boguraev (2002). Automatic Glossary Extraction: Beyond Terminology Identification. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei.
- Pazienza, M. T., M. Pennacchiotti, and F. M. Zanzotto (2005). Terminology extraction: an analysis of linguistic and statistical approaches. *Knowledge Mining* 185, 255–279.
- Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1-2), 137–158.
- Pecina, P. and P. Schlesinger (2006). Combining Association Measures for Collocation Extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006)*, Sydney, pp. 651–658.
- Pedersen, T. (1996). Fishing for Exactness. In *Proceedings of the South-Central SAS Users Group Conference (SCSUG-96)*, Austin, TX.
- Purandare, A. (2004). *Unsupervised Word Sense Discrimination by Clustering Similar Contexts*. Msc thesis, University of Minnesota.
- QasemiZadeh, B., P. Buitelaar, T. Chen, and G. Bordea (2012). Semi-Supervised Technical Term Tagging With Minimal User Feedback. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, Istanbul, pp. 617–621.
- QasemiZadeh, B. and S. Handschuh (2014). The ACL RD-TEC: A Dataset for Benchmarking Terminology Extraction and Classification in Computational Linguistics. In *Proceedings of the 4th International Workshop on Computational Terminology*, Dublin, pp. 52–63.
- Sclano, F. and P. Velardi (2007). TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. In *Proceedings of the 9th Conference on Terminology and Artificial Intelligence (TIA 2007)*, pp. 8–9.
- Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28(1), 11–21.
- Warburton, K. (2013). Processing terminology for the translation pipeline. *Terminology* 19(1), 93–111.
- Zhang, Z., J. Iria, C. Brewster, and F. Ciravegna (2008). A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, pp. 2108–2113.

Acquiring Verb Frames for a Text Simplification Lexicon in the Medical Domain

Ornella Wandji Tchami

¹ STL-UMR8163-University of Lille3, France

² IWIST-University of Hildesheim, Germany
`wandji@uni-hildesheim.de`

Abstract. In this paper, we present a method for the acquisition of medical frames of verbs from a medical corpus, with the purpose of creating a text simplification resource which aligns verb constructions from medical expert language with their lay equivalents. Our approach uses a syntactico-semantic verb classification, a medical terminology and three subcorpora differentiated according to the level of expertise of their readership. The evaluation gives a precision that varies between 0.50 and 0.87 according to the verbs, and shows that the quality of the results depends on the quality of the semantic annotation.

Keywords: medical corpora; doctor vs patient communication; text simplification resource; specialised verbs; argument structure; verb frames; verb semantic features; specialised corpora.

1 Context and Goal of the Research

Standard medical language is sometimes hard to understand for non-expert users [18], due to linguistic complexity [20] and to the use of domain-specific vocabulary. These issues can cause problems in the communication between medical experts and patients [14].

To overcome this communication issue, researchers in Natural Language Processing suggest the use of text simplification as a means to ease the understanding of the medical practitioners' language by lay people [28, 4]. Text simplification can be defined as the process of reducing the linguistic complexity of a text, while still retaining its original information and meaning [23]. The main goal of simplification is to make information more accessible to the targeted audience. Simplification may cover the syntax [3], the lexicon [10, 1, 2, 8, 16], or simply focus on surface characteristics of the text, e.g. the number of characters and syllables per word, capitalization, and punctuation [15]. Several researchers have investigated the use of text simplification as a means for facilitating access to medical texts, by simplifying terminology [10, 13]. These studies demonstrate that any simplification requires resources, which presupposes the description of the specificities of the language that needs to be simplified.

This work is part of a research project that aims at creating a text simplification resource for specialised French medical texts. Our method therefore contributes to research on lexical simplification. Much research work addresses this topic; however, most of it focuses on the English language while very little concerns French. More importantly, our method differs from the others, as we deal with verbs and their arguments rather than nominal entities only.

Indeed, the resulting resource is expected to contain medical experts verbal constructions (called *frames*³), aligned with their lay equivalents. The texts used in the whole project, as well as in the present paper, were analysed in our previous studies which describe the similarities and specificities of medical texts with different levels of specialisation [27, 26, 25]. The results of these experiments showed that experts medical texts are characterised by specialised verb frames, which are specific to the experts language. However, they also showed that lay texts could provide quasi synonymous verb frames, more common to the non-experts, which can be used as equivalent substitutes for the specialised verb frames, in the framework of text simplification. Some examples of verb frames⁴ are presented in Table 1:

Table 1. Expert verb frames with their corresponding lay equivalents

Expert (specialised) verb usages	Lay equivalent usages
<i>Le S subit/développe/relève (d')une D</i>	<i>Le S a/fait/souffre d' une D</i>
<i>The S develops/suffers (from) a D</i>	<i>The S has a D</i>
<i>Le J diagnostique le S</i>	<i>Le J dépiste le S</i>
<i>The J diagnoses the S</i>	<i>The J screens the S</i>

Our aim is to identify the specialised frames of verbs (left column), for the purpose of simplification. For future work, our objective will be to align them with their corresponding lay equivalents (right column).

The purpose of this study is therefore to propose a method for the acquisition of specialised medical frames of verbs. The applied method requires: a lexical resource of verbs [12] for the acquisition of syntactic patterns and the selection of specialised medical frames of verbs; a medical terminology [5], which is used to add semantic information to the syntactic patterns of verbs ; and a corpus composed of three subcorpora differentiated according to the level of expertise of their author and intended readership, which is used for the acquisition of sentences illustrating the frames.

Our approach is influenced by Frame Semantics [11], a theory whose basic idea is that the meaning of a lexical unit can be best understood on the basis of a semantic frame, *i.e.* a conceptual scenario which describes a type of event, relation, or entity and the participants in it. The principles of Frame Semantics are implemented in the FrameNet project [21]. Some research work

³ In our study the word *frame* refers to a subcategorisation scheme in which the arguments are associated to semantic categories provided by the Snomed terminology.

⁴ In these examples: *S*= *patient*, *J*=*doctor*, *D*=*disease*, *C*=*medication*.

has applied Frame Semantics to specialised languages [6, 22, 19, 17], but none of these studies was carried out in the framework of text simplification. We adapt and use a FrameNet-like data modeling for the purpose of text simplification. However, our approach differs to some extent from FrameNet’s. Firstly, we propose a bottom-up approach (from text to frames), while FrameNet is top-down. Secondly, our semantic annotation of verb arguments is based on semantic categories provided by the Snomed medical terminology. These categories describe semantic sorts (*disease, anatomy, chemical product, etc.*) of arguments rather than their semantic roles (*healer, patient, etc.*)

2 Material

2.1 Corpus

The corpus is composed of three different medical subcorpora of written French. As described in Table 2, these subcorpora contain non-aligned texts from different medical subdomains, which are distinguished according to the level of expertise of their authors and of the intended audiences. They are collected from the *CISMeF*⁵ portal, which indexes medical texts according to their targeted audience: medical experts, medical students and general public.

Table 2. Size of the subcorpora

Subcorpus	Size (words)	Description
expert to expert texts	1,785,665	scientific publications, reports
expert to student texts	1,755,497	didactic supports for students
expert to lay texts	1,627,466	documentation, brochures

The subcorpora have similar sizes, more than one million words each (Table 2). They are used for the extraction of semantic frames of the verbs. The diversity of the subcorpora plays an important role in this study because it will make it possible to observe the frequency of the extracted frames in the different types of texts.

2.2 Lexical Resource of Verbs

The proposed approach is based on data (example sentences) taken from a dictionary of French verbs called *Dictionnaire Électronique des Verbes Français* (LVF) [9], which is an electronic version of *Les Verbes français*, a classification of French verbs made by Jean Dubois and Françoise Dubois-Charlier [12]. This resource, with 25 610 entries, proposes a syntactico-semantic classification of verbs, based on their valency patterns. For each entry, the LVF database provides different types of information: semantic class, valency pattern (intransitive, transitive,

⁵ <http://www.cismef.org/>

pronominal, etc.), meaning (a synonym, a definition, or an explanation), domain and example, etc. In this study, we have only considered the verb usages that are marked in LVF as belonging to the medical domain, a total of 318 items. The example is usually a simple sentence illustrating the valency pattern and the semantic class of the entry word:

(1) *Le médecin admet un malade dans ce service.*

The sentence in (1) describes the features of *admettre*, which is a transitive verb requiring a direct object, plus a prepositional complement either introduced by *dans* or *de*. The words which occupy the arguments positions in the LVF example sentences (*médecin*, *malade*, *service*) stand for semantic sorts and fulfil the selectional restrictions of the verb arguments. We exploit this in a mapping with the Snomed ontology (see section 3.3).

2.3 Semantic Resource

We use the *Snomed International Terminology* [5], a term base which groups medical terms into eleven semantic categories, of which nine are considered in this work⁶:

T: Topography or anatomical locations (*e.g.*, *coeur*, *cardiaque*, *bras*, *vaisseau*);
S: Social status (*e.g.*, *mari*, *soeur*, *mère*, *ancien fumeur*, *donneur*);
P: Procedures (*e.g.*, *césarienne*, *transducteur ultrasons*, *télé-expertise*);
L: Living organisms, such as bacteria and viruses (*e.g.*, *Bacillus*, *Salmonella*);
plants (*e.g.*, *fougère*, *pomme de terre*), but also animals (*e.g.*, *singe*, *chien*);
J: Professional occupations (*e.g.*, *équipe de SAMU*, *anesthésiste*, *assureur*);
F: Functions of the organism (*e.g.*, *pression artérielle*, *détresse*, *insuffisance*);
D: Disorders and pathologies (*e.g.*, *obésité*, *hypertension artérielle*, *cancer*);
C: Chemical products and food (*e.g.*, *médicament*, *héparine*, *bleu de méthylène*);
A: Physical agents and artefacts (*e.g.*, *cathéter*, *prothèse*, *tube*).

The Snomed International Terminology was chosen because it is one of the largest medical terminologies freely accessible for French. This terminology was defined by medical experts, in order to improve patient care through the development of systems to record health care encounters accurately. However, we use the Snomed categories for linguistic purposes: they are used as semantic labels for the semantic annotation of arguments. The original version of Snomed contains 144 267 entries (mainly French nouns, noun phrases and adjectives). These entries may not necessarily cover all domain notions in our texts. For this reason, we used a version of Snomed that was enriched in relation with the corpus used [24].

3 Method

We aim at acquiring medical frames of verbs, using a syntactico-semantic verb classification and the Snomed medical terminology. These frames will serve as

⁶ Two semantic classes containing modifiers are not taken into consideration.

material for the creation of a text simplification resource. The various steps of our method are described in Fig. 1. The light grey box represents the aim of this work, while the dark grey boxes signal interactive tasks which require manual work.

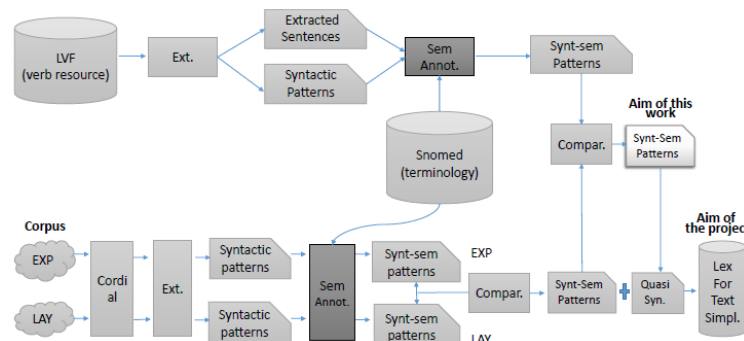


Fig. 1. General schema of the method

3.1 Corpus Pre-processing and Selection of Verbs

The subcorpora are collected, converted into plain text, recoded in UTF-8 and cleaned up to be easily processable. A dependency syntactic analysis is then performed with the *Cordial* dependency parser [7], which makes it possible to identify and retrieve the verbs and their arguments. The experiment described in this paper concerns 11 verbs, selected as follows: (1) being part of the LVF’s entries, (2) having at least 30 occurrences in the corpus (cumulated frequency).

3.2 Semi-Automatic Acquisition of Frames from the Corpus

The extraction of medical frames from the corpus is performed with an interactive system (named *FrameExtSystem*) that was implemented in a previous study [24]. As input, the *FrameExtSystem* takes different medical texts that have been parsed beforehand with the *Cordial* parser; it performs a semantic annotation of the parsed texts, using the Snomed terminology: by assigning Snomed categories (see section 2.3) to nominals in each sentence. The result of the semantic annotation is manually improved, in order to process unlabelled terms and to correct erroneous labels. Afterwards, the corrected output of the semantic annotation is given back to the system, which finally extracts the verb frames from the various texts. The output is a file where each verb is associated to frames extracted from each subcorpus, with the corresponding frequency.

3.3 Semantic Annotation of Sentences and Acquisition of Frames from the LVF

Example sentences are automatically extracted from the LVF entries which are marked as belonging to the medical domain. These sentences illustrate different valency patterns of the verbs (see section 2.2).

(2) *Le médecin administre un remède au malade*
⇒ *s_J administrer cod_C coi_S*

As illustrated in example (2), not only are the arguments positions are instantiated by words that fulfil the selectional restrictions of the verb, but in addition, these words are usually generic medical terms contained in Snomed, that allow us to perform the semantic labelling of arguments, according to the technology used in section 3.2.

3.4 Comparison of Frames and Selection of Entries for the Text Simplification Resource

We compare the frames acquired from LVF and those extracted from our sub-corpora in different ways : (i) qualitatively, in terms of corpus coverage of LVF (corpus frames covered in and those absent from the LVF; LVF frames attested in the corpus vs. absent from it), and (ii) quantitatively, between our three sub-corpora, to identify preferences in the verb use in relation with the expert vs. lay texts. On the basis of these figures, we select verb readings whose frames from the expert texts can be replaced in the text simplification process with more understandable frames of verbs frequent in texts written for lay people. This interactive process relies on linguistic and terminological features of the verbs and their arguments provided by the LVF and the Snomed terminology. For more reliability, the alignment of expert-lay pairs will be done in collaboration with medical experts, and we are planning to perform an evaluation of our final results. The results of the first two steps above are presented in the following section.

4 Results and Discussion

4.1 Sentence Extraction and Automatic Acquisition of LVF Frames

From LVF, 318 sentences with full noun phrases as subjects and complements are extracted. From these, 420 medical frames are derived.

Some sentences contain two constructions separated by a full stop (*Le médicament n'apaise plus sur P. L'aspirine apaise dans ce cas.*), or two arguments separated by a comma (*On apaise un malade, la douleur.*). Such sentences were split into two, hence generating two distinct frames from a single LVF example sentence. In certain cases like example (3), these frames were semantically different, leading to polysemy:

(3) *On apaise un malade, la douleur*

a) *On apaise un malade* \Rightarrow *On apaise un S* (Someone relieves/appeases the patient)

b) *On apaise la douleur* \Rightarrow *On apaise F* (Someone alleviates/eases the pain)

Our method hence helps to detect verb polysemy (thanks to the Snomed semantic categories), even when it is described in a single LVF example sentence.

4.2 Comparison of Frames and Selection of Entries for the Text Simplification Resource

Table 3 provides results of the comparison between the frames generated from the LVF sentences and those extracted from the corpus, for the 11 verbs selected as examples to illustrate our work. For each verb, the table shows the number of LVF frames (*LVF_fram*) ; the number of common frames between the LVF and the corpus (*Com_fram*), plus their number of occurrences in the corpus (*Occ com_fram*) ; the number of corpus-specific frames (types): *Corpus_fram*, plus their number of occurrences in the corpus (tokens): *Occ corpus_fram*.

Table 3. Number of frames per verb in the corpus and in the LVF

Verb	LVF fram	Nb com_fram	Occ. com_fram	Corpus fram	Occ. corpus fram
abaisser	1	0	0	19	55
admettre	1	1	1	33	74
diagnostiquer	1	1	2	30	101
imposer	1	0	0	66	388
indiquer	1	1	3	193	768
relever	1	1	1	78	183
subir	1	1	4	43	92
suivre	1	0	0	142	488
survivre	2	1	3	14	31
traiter	1	1	1	107	297
stimuler	1	1	5	39	76

Table 3 shows that the number of corpus frames (types of frames) per verb can be very high. This is because several frames have syntactic variants (up to 4 for certain frames), coupled with the fine semantic granularity of the Snomed categories (many possible combinations between the 9 Snomed categories and the 3 syntactic positions). The second remark is that almost all the LVF frames are found in the corpus, most of the time with different variants, e.g. active vs. passive voice. In addition, the numbers in Table 3 show that our corpus still has much to offer as medical frames of verbs are concerned. Indeed, the minimum number of corpus-specific frames for a single verb is 14, versus 1 for the LVF. This is a consequence of the syntactic variation of frames.

Table 4 presents the 8 LVF frames attested in the corpus, for the 11 sample verbs. Before interpreting the results, it is important to underline that some frames have a low token frequency, because they have other variants in the corpus. For example, *s_J diagnostiquer cod_D* has three variants: *s_J diagnostiquer*

Table 4. Common frames with their frequencies from the three subcorpora

Verb frames	Corpus	Exp	Stu	Lay
s_J admettre cod_S coi_S	1	1	0	0
s_J diagnostiquer cod_D	2	2	1	13
s_T indiquer cod_F	3	1	2	0
s_On relever coi_D	2	1	1	0
s_F stimuler cod_F	5	1	2	2
s_On survivre coi_D	3	1	1	1
s_On subir cod_P	64	4	17	17
s_J traiter cod_S	1	0	1	0

cod_D coi_S, and its passive forms *s_D être-diagnostiqué par coi_S*, and *s_D être-diagnostiqué chez S* (passive form with an omitted agent).

Table 4 also shows that some of the common frames are frequent in the corpus while others are not. The variation of the frequency curve between the subcorpora (*e.g. s_On subir cod_P*) functions as an indicator of corpus-specific usages of verbs. Indeed, the high frequency of a frame in the expert subcorpus signals a specialised meaning of the verb which belongs to the experts vocabulary. A high frame frequency in the lay subcorpus signals a verb usage that is more common to lay people (*s_J diagnostiquer cod_D*). Low frequency frames are not disregarded. Instead, they can be as significant as high frequency frames, because they sometimes represent highly specialised usages of the verb. A good example is the frame *s_S relever coi_D*, where *relever* means *to suffer from*. This frame is present only twice in the corpus, namely in the expert and student subcorpora:

(4) *L' exonération du ticket modérateur peut être donnée [...] lorsque le patient relève d' une affection de longue durée.*

The frames presented in Table 4 and selected thanks to the LVF resource hence constitute potential candidates for the text simplification resource. These results show how relevant the use of the LVF can be in the process of selecting frame candidates for the creation of our text simplification resource. However, the corpus provides many more frames that are not found in the LVF but that can be exploited for the text simplification resource. Table 5 gives some examples of these corpus-specific frames:

Table 5. Examples of corpus-specific frames with their frequencies

Frames	Freq	Frames	Freq	Frames	Freq
s_F est abaissé	21	s_D imposer cod_P	49	s_S subir cod_D	16
s_S est admis coi_S	20	s_P imposer cod_P	38	s suivre cod_P	42
s_D est diagnostiqué	41	s_F est indiqué	27	s suivre cod_F	30
s_D diagnostiquer coi_S	11	s_P est indiqué	73	s_On traiter cod_D	30
s_P est imposé	94	s_P indiquer coi_D	16	s_D est traité	23

The 15 frames presented in Table 5 were selected on the basis of their frequency in the corpus (minimum 10). However, low frequency frames can also be relevant. For example, *s_J diagnostiquer s_S* is a frame which was found only in the expert subcorpus, and with less than 10 occurrences. In this context, the

verb is synonymous with *dépister* (to detect/discover), which is an unusual synonymy for the lay persons ([...] *il faut entreprendre la médication contre le TDAH à la recommandation de la personne qui diagnostique et suit le patient ayant le TDAH*). These 15 frames in table 5 represent less than half of the corpus-specific frames, which shows how rich the corpus data are, and how necessary it is to consider these frames when gathering the material for the text simplification resource.

5 Evaluation of the Method

We propose a two stage (individual and group) semi-automatic evaluation of the applied method. For each verb, the evaluation is based on the following information recorded in Table 6 : number of frames generated from the LVF (*LVF*), number of frames found only in the corpus (*Corpus*) (these corpus-specific frames are evaluated as follows: number of frames fully annotated with semantic information from Snomed (*full*), number of partially annotated frames (*part*), number of erroneous frames (*error*)), number of common frames between the corpus and the LVF (*Com*), precision of fully annotated frames (*precision full*). The erroneous frames were identified manually, while all other information was obtained automatically. For each verb, the precision is the ratio between the number of full frames and the total number of frames of the verb.

Table 6. Results of the evaluation

Verbs	Nb frames Corpus				LVF	Com	Precision full
	full	part	error	total			
abaisser	12	5	2	19	1	0	0.63
admettre	20	12	2	34	1	1	0.58
diagnostiquer	23	6	2	31	1	1	0.74
imposer	38	22	6	66	1	0	0.57
indiquer	110	82	2	194	1	1	0.56
relever	57	20	2	79	1	1	0.72
subir	33	9	3	45	1	1	0.73
suivre	74	68	0	142	1	0	0.52
survivre	8	6	0	14	2	1	0.57
stimuler	35	3	2	40	1	1	0.87
traher	55	52	2	109	1	1	0.50
Total	465	285	23	773	11	8	0.601

According to the evaluation results recorded in Table 6, the precision varies depending on the verbs. It goes from 0.50 for *traher* to 0.87 for *stimuler*, resulting in an average precision of 0.60. This value is highly dependent on the quality of the semantic annotation of the subcorpora. Verbs (*traher*, *suivre*, *indiquer*) with the lowest precision are those whose semantic annotation was not completely enhanced manually (due to time constraint), while sentences with a totally improved semantic annotation show better scores. The high number of corpus-specific frames obtained for the 11 analysed verbs shows the extent to which our method can be productive for the acquisition of medical frames of verbs from medical corpora. These frames could be used for extending the content of the LVF.

6 Conclusion and Future Work

In this paper, we have described and evaluated a method for the acquisition of medical frames of verbs, for the creation of a text simplification resource. The method uses sub-corpora differentiated according to the level of expertise of their author and intended audience, a syntactico-semantic classification of verbs and an existing medical terminology. Our approach allows us to identify and select expert candidate verb frames from the corpus, that will be further aligned with their equivalents, for the purpose of text simplification. The evaluation shows that the precision (between 0.50 and 0.87) depends on the quality of the semantic annotation, which means that a better semantic annotation would improve the performance of the system. For future work, we are planning to improve the semantic annotation for better results. We would also like to further exploit the LVF sections dedicated to other domains which are related to the medical domain, *e.g. anatomy, pathology*.

References

1. Biran, O., Brody, S., Elhadad, N.: Putting it simply: a context-aware approach to lexical simplification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, Vol 2. pp. 496–501. Association for Computational Linguistics (2011)
2. Bott, S., Rello, L., Drndarevic, B., Saggion, H.: Can Spanish be simpler? lexis: Lexical simplification for Spanish. In: CoLing. pp. 357–374 (2012)
3. Brouwers, L., Bernhard, D., Ligozat, A.L., Franois, T.: Syntactic sentence simplification for French. In: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL. pp. 47–56 (2014)
4. Chmielik, J., Grabar, N.: Détection de la spécialisation scientifique et technique des documents biomédicaux grâce aux informations morphologiques. TAL 51(2), 151–179 (2011)
5. Côté, R.A.: Répertoire d’anatomopathologie de la SNOMED internationale, v3.4. Université de Sherbrooke, Sherbrooke, Québec (1996)
6. Dolbey, A., Ellsworth, M., Scheffczyk, J.: BioFrameNet: A domain-specific FrameNet extension with links to biomedical ontologies. In: KR-MED (2006), 87–94
7. Dominique, L., Nègre, S., Séguéla, P.: L’ analyseur syntaxique Cordial dans Passage. Actes de TALN 9 (2009)
8. Drndarević, B., Saggion, H.: Towards automatic lexical simplification in Spanish: an empirical study. In: Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations. pp. 8–16. Association for Computational Linguistics (2012)
9. Dubois, J.: Dictionnaire électronique des verbes français. Linx 3(1), 213–230 (1991)
10. Elhadad, N.: Comprehending technical texts: predicting and defining unfamiliar terms. In: AMIA. pp. 239–243 (2006)
11. Fillmore, C.: Frame semantics. Linguistics in the morning calm pp. 111–137 (1982)
12. François, J., Le Pesant, D., Leeman, D.: Présentation de la classification des verbes français de Jean Dubois et Françoise Dubois-Charlier. Langue française 153(1), 3–19 (2007)
13. Grabar, N., Hamon, T.: Automatic extraction of layman names for technical medical terms. In: Healthcare Informatics (ICHI). pp. 310–319. IEEE (2014)

14. Jucks, R., Bromme, R.: Choice of words in doctor-patient communication: an analysis of health-related internet sites. *Health Commun* 21(3), 267–77 (2007)
15. Kanungo, T., Orr, D.: Predicting the readability of short web summaries. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. pp. 202–211. ACM (2009)
16. Leroy, G., Endicott, J.E., Mouradi, O., Kauchak, D., Just, M.: Improving perceived and actual text difficulty for health information consumers using semi-automated methods. In: *AMIA* (2012)
17. L’Homme, M.: Adding syntactico-semantic information to specialized dictionaries: an application of the FrameNet methodology. *Lexicographica* 28, 233–252 (2012)
18. McCray, A.: Promoting health literacy. *J of Am Med Infor Ass* 12, 152–163 (2005)
19. Pimentel, J.: Description de verbes juridiques au moyen de la sémantique des cadres. In: *TOTH* (2011)
20. Putz, M.: Approaching linguistic complexity in medical care. *International Journal of Anthropology* 23(3-4), 275–284 (2008)
21. Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C.R., Scheffczyk, J.: *FrameNet ii: Extended theory and practice*. Tech. rep., FrameNet (2006), available online <http://framenet.icsi.berkeley.edu>
22. Schmidt, T.: The Kicktionary. A Multilingual Lexical Resource of Football Language, pp. 101–134 (2009)
23. Siddharthan, A.: Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In: *EACL*. pp. 722–731 (2014)
24. Wandji Tchami, O., Grabar, N.: Towards automatic distinction between specialized and non-specialized occurrences of verbs in medical corpora. In: *Proceedings of Computerm*. pp. 114–124. Dublin, Ireland (August 2014)
25. Wandji Tchami, O., Grabar, N., Heid, U.: Syntagmatic behaviors of verbs in medical texts: Expert communication vs. forums of patients. In: *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence*. pp. 99–106. Universidad de Granada, Granada, Spain (November 2015)
26. Wandji Tchami, O., L’Homme, M., Grabar, N.: Frame semantics-based study of verbs across medical genres. In: *e-Health - For Continuity of Care - Proceedings of MIE2014, the 25th European Medical Informatics Conference, Istanbul, Turkey, August 31 - September 3, 2014*. pp. 1075–1079 (2014)
27. Wandji Tchami, O., L’Homme, M., Grabar, N.: Discovering semantic frames for a contrastive study of verbs in medical corpora. In: *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence*. Villetaneuse (2013)
28. Zeng-Treiler, Q., Tse, T.: Exploring and developing consumer health vocabularies. *J of Am Med Infor Ass* 13, 24–29 (2006)

Design of a corpus for mathematical knowledge transfer to 6 – 12 year-old children

María Pozzi

Centro de Estudios Lingüísticos y Literarios, El Colegio de México, Camino al Ajusco 20,
Pedregal de Santa Teresa, 10740, Mexico City, Mexico
pozzi@colmex.mx

Abstract. This paper presents the design of an annotated corpus intended as a tool to help children understand the basic mathematical concepts learned during the six years of primary school education in Mexico. It contains all nine official maths textbooks. Annotation is organised into two levels each one of them having several layers. The first one includes metadata, lexical and terminological information and the second includes additional semantic/conceptual and practical information. It is structured in two XML files: the first file contains the corpus itself annotated with tags corresponding to metadata, lexical, and terminological analyses, and the second file contains semantic/conceptual and practical information.

Keywords: corpus-based terminology, knowledge transfer, mathematics terminology, children's vocabulary

1 Introduction

Traditionally, one of the main differences between terms and general language words is that terms are explicitly learned and not picked-up from everyday life experience. Cabré (Cabré 1993, 222-3) points out the main features that differentiate words from terms which include: a) terms have a referential function, b) terms are used in one or more specific domains of knowledge, c) terms are used in formal communicative situations, d) terms are used in a professional or scientific discourse, and e) terms are used by specialists. For terminography, on the other hand, ISO 704:2009 states (ISO 704 2009, 22) that “a terminology shall not include terms that are so general that can be thought of as general language words and are adequately defined in general language dictionaries”.

As a general rule, I would agree with these statements. However, how should the first scientific and mathematical terms that children learn at school be considered? Are these terms or general language words? The answer depends on who is posing the question. For an educated adult or a specialist these are probably general language words, but for children who have to learn them, teachers that have to teach them, textbook writers, education policy makers and others working in related fields, these should obviously be considered as terms, since they satisfy all but one of Cabré's criteria for being terms (children are no specialists). Because of a lack of consensus on their nature

or perhaps because these units share most of their characteristics with both general language words and terms, the explicit science and maths vocabularies learned by children during the first six years of primary school are in no man's land and therefore have not received a lot of attention from terminologists and linguists alike. For the purpose of this project, the linguistic units representing mathematical concepts at the lowest level of specialisation are assumed to be terms.

But one thing is for certain: from that age on, children will eagerly learn and like maths depending on how well they understand and make their own these first concepts taught at school. Thus the great importance of acquiring both concept and terms right from the time of their first encounter.

This paper presents some features of the Corpus of Primary School Maths Texts (COPSMAT)¹, a subset of the Corpus of basic scientific texts in Mexican Spanish, (COCIEM), a multi-purpose infrastructure developed to identify and characterise the basic scientific vocabulary in Mexican Spanish. At this stage, one of COPSMAT's aims is to have a tool for identifying, analysing and studying the maths vocabulary learned at school. Ultimately, it has been conceived to help primary school children —and teachers— to strengthen their understanding of basic mathematical concepts and for the appropriate use of the corresponding terms.

2 Background

As most western countries, Mexico has replaced its educational system from a traditional to a constructivist approach, mainly in maths and science, at the primary school level (6 to 12 year-olds). This means that following Piaget's (Piaget 1967) ideas, children are expected to discover principles and construct hands-on their own understanding of concepts based on observation and on 'doing things'. Vygotski (Vygotski 1978), went further by stating that knowledge is constructed in a social context before it is actually acquired. This change of approach has had a considerable effect on the teacher-children relationship in the classroom and on the content and presentation of textbooks as well. Teachers, in principle, are supposed to provide the appropriate guidance and environment to students so that they can make their own observations and produce their own mental constructions.

The change in Mexico, although welcome in principle, has not proved effective for several reasons, among which, in my opinion, the following three stand out:

- Teachers were not offered compulsory training on the constructivist approach to teaching and learning. After twenty years, many are still oblivious to these changes.
- Although textbooks follow the constructivist approach, most schools do not carry out teacher assessment to ensure they follow this approach. As a consequence, there is an obvious incompatibility between teaching methods and textbooks.
- Maths textbooks contain activities, exercises, illustrations, problems, etc., but almost no explicit conceptual information, definitions or even explanations of concepts. As a result, if children do not understand what they are doing, they cannot conceptualise

¹ *Corpus de textos de matemáticas de primaria en español de México (COTEMP).*

well and in turn, the following concepts they are supposed to learn will not be understood either. And this process can go on and on...

Perhaps this lack of conceptual understanding could be one of the key factors for the result of the Programme for International Student Assessment (PISA) conducted by the Organisation for Economic Co-operation and Development (OECD) that measures the level of achievement in mathematics and science in children aged 6-15 years (OECD 2009). According to the report for Mexico (OECD 2008): "In Mexico [...] the mean score in PISA performance in science and mathematics is well below average". From the 48th place it took in 2008, it dropped to 53rd place out of the 65 countries studied in 2012 (OECD 2013). Furthermore, out of the thirty-four OECD member states, Mexico occupies the last place both in maths and science performance. These are alarming results.

With this background in mind, I believe we are in position to help children to conceptualise and to acquire those very first mathematical concepts —and terms— that are the basis on which they will construct their own mathematical knowledge in years to come, by strengthening their concept understanding once they have made initial observations and have had hands-on practice on a particular concept or set of related concepts. For this purpose, we decided to prepare a corpus-based terminographical product providing conceptual, linguistic and practical information as well as examples, illustrations, exercises and other activities. It is intended to complement —not substitute— the information provided in maths textbooks.

3 Corpus of Primary School Maths Texts (COPSMAT)

COPSMAT contains all nine official maths textbooks used in primary schools in Mexico. In order to be able to identify as exhaustively as possible all maths terms taught and learned in primary school, each complete textbook was included, except for figures and tables. The total number of graphic words (tokens) is 125,142 while the number of types amounts to 8,533.

It is an XML file with three layers of annotation²:

1. metadata: bibliographic reference, area to which the textbook belongs and school year in which the textbook is used;
2. lexical: for each word, its lemma and POS;
3. terminological: for each identified term, its lemma and POS.

Since COPSMAT is a subset of the Corpus of basic scientific texts in Mexican Spanish (COCIEM), the lexical markup was already in place. In turn, COCIEM was tagged using the data provided by the Corpus del español de México contemporáneo (Lara et al. 1979).

² These are the original three layers of annotation already marked when we created this subset of COCIEM to generate COPSMAT. We have added more layers and an additional level of annotation with 2 layers.

Most of the terms occurring in COPS MAT had already been identified and tagged as a result of the term extraction and validation processes applied to COCIEM. These processes are fully described in Cabrera-Diego et al. (Cabrera-Diego et al. 2011) and Vivaldi et al. (Vivaldi et al. 2012).

3.1 Further term extraction

The first practical problem we faced was that since most of the very basic maths terms learned in the first years of primary school are thought of as general language words (e.g. line, distance, time, area, plus, unit, sequence), these were not included in the original list of term candidates. But since we already had the POS tag for all single words contained in COPS MAT, the next step towards the exhaustive identification and validation of maths terms was to generate left and right concordances for all nouns, adjectives and verbs using WordSmith Tools 5.0 (Scott 1996) to get an additional list of single-word and multiword term candidates which had yet to be validated.

3.2 Term validation

To validate the new term candidates, we firstly applied the process developed by Cabrera-Diego et al. (Cabrera-Diego et al. 2011) and Vivaldi et al. (Vivaldi et al. 2012), in which Wikipedia is used to determine whether a term candidate belongs to a defined specific domain, in this case, mathematics. It is well known that Wikipedia is organised into two connected graphs, the *category graph* and the *page graph*. In turn, the category graph is organized as a taxonomy where each category may be connected to an arbitrary number of super or sub categories. Wikipedia articles, on the other hand, are linked between them forming a directed graph. Both graphs are connected together since every article is assigned to one or more Wikipedia categories. The following procedure was then applied: a) the domains of interest were defined as mathematics, geometry and statistics; b) for each term candidate, find a Wikipedia page; c) find all Wikipedia categories associated to that page; d) explore the category graph in a recursive manner to follow up all super category links until one of the defined domains or the Wikipedia top category is reached. After some additional procedures and calculations, if one of the categories found coincides with the defined domains of interest, the term candidate is validated. Figure 1, taken from Vivaldi et al. (Vivaldi et al 2012, 3823), shows an example of this process.

The second part of the validation process, applied to those term candidates that for whatever reason failed to be validated, was carried out manually by a mathematician.

The final list of validated maths terms learned in primary school contains 1097 terms. At this stage all designations including synonyms, abbreviations, symbols and signs are taken to be terms representing a concept. For multi-word terms, the corresponding POS was manually tagged.

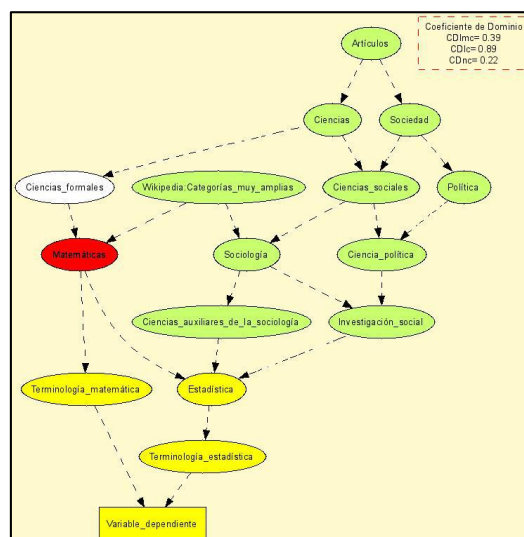


Fig. 1. Graph for term candidate *variable dependiente*

4 Additional information

To complement COPSMAT, new annotation tags were added to include further linguistic/terminological analyses, semantic/conceptual and practical information. The annotation corresponding to linguistic and terminological information was stored in the corpus file itself, while the semantic/conceptual and practical information was stored in a separate file. This was done in such way as to optimise the use of computer storage since it is textual information that would have to be repeated for each occurrence of the term in question, so the corpus file stores a pointer to the second file where the information is stored only once.

4.1 Linguistic/terminological annotation

To achieve COPSMAT's objective, more detailed terminological information had to be specified:

- preferred term, in the case where two or more terms designate one concept.

gráfica de columnas → gráfica de barras <sg tt="TP" l="gráfica de barras">gráfica de columnas</sg>

- abbreviation (with the associated full form)

km → kilómetro <g tt="abr" l="kilómetro">km</g>

- symbol/sign (with the associated full form)

$\infty \rightarrow$ infinito `<g tt="sim" l="infinito"> ∞ </g>`

- synonym (with the preferred term)

sistema decimal \rightarrow sistema de números decimales `<sg tt="TP" l="sistema de números decimales">sis-tema decimal</sg>`

- homonym

mediana \rightarrow geometría `<g re="hom" l="geometría">mediana</g>`
 mediana \rightarrow estadística `<g re="hom" l="estadística">mediana</g>`

- antonym

cóncavo \rightarrow convexo `<g re="ant" l="convexo">cóncavo</g>`

4.2 Semantic/conceptual annotation

Semantic information is crucial for a better understanding of the underlying concepts and so is the presentation of this information. It includes

- *subdomain* (arithmetic, arithmetic operations, geometry, statistics and graphs and numbers and systems of numbers). For example,

círculo – geometría

- *encyclopaedic information* which might be useful for a better understanding of the concept in question. This information does not form part of the definition. For example,

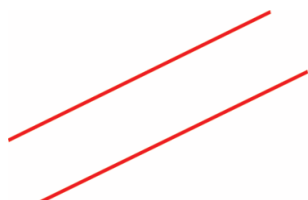
circunferencia	<ul style="list-style-type: none"> • perímetro del círculo • todos los puntos de la circunferencia están a la misma distancia del centro del círculo
----------------	--

- *example* illustrating the concept itself rather than the use of the term. For example,

circunferencia	<ul style="list-style-type: none"> • una llanta, un aro, un anillo son ejemplos de una circunferencia
----------------	--

- *illustration*, in which one or more drawings are provided to illustrate the concept or its usage. Whenever possible, an illustration of the abstract concept and one of an application were provided:

líneas paralelas



- *related terms / cross references*, where a list of terms whose underlying related concepts is provided:

denominador	fracción numerador común denominador
masa	peso
peso	masa

- *definition*

To ensure that the definitional model to be used was the one that helped children to understand the concept better, we carried out an experiment involving both, teachers and children of different schools and age (9 and 12 year-olds) to detect the form which made the acquisition of the concept easier for them. It is important to point out that for many children this was their first encounter with definitions of mathematical concepts although in principle, they should have known/understood the underlying concept they were dealing with. The reality, in many cases, was very different. We presented them with twenty concepts together with three different definition models and they had to select the order in which it was easier for them to understand the definitions and why.

The first option presented a classical terminological definition (superordinate concept followed by essential characteristics) written in words they already knew. For example,

número cardinal	número que expresa unidades de una serie, como uno, dos, cien, etc., que representa una cantidad pero no orden y que se usa para contar y hacer operaciones aritméticas.
-----------------	--

A second option introduced the “definition” in a context, in the style of the Collins Cobuild Dictionary. For example,

equivalente	Si dos o más cantidades tienen el mismo valor, aunque estén expresadas de distinta manera, son equivalentes.
-------------	--

The third option consisted of a list of individual characteristics. For example:

gráfica circular	<ul style="list-style-type: none"> — gráfica que tiene la forma de un círculo — el círculo está dividido en sectores de diferente tamaño — cada sector representa una cantidad o el porcentaje del total de los datos
------------------	--

The result of this experiment was clear, children felt more comfortable and understood the definition better when it was presented as a list of characteristics closely followed by those presented in context. Younger children had some difficulty understanding definitions presented in the classical model. Therefore, definitions were drafted as a list of characteristics. This decision had the unforeseen advantage of being able to investigate amongst other things, the relationship between concepts sharing characteristics.

4.3 Practical information annotation

The practical information annotation provides an important link between more abstract concepts and how they are applied in everyday life. By adding it, it is hoped these concepts will become meaningful to children of this age group.

- *exercise*, explained step by step:

resta	<ul style="list-style-type: none"> • Neil Armstrong fue el primer hombre que caminó en la Luna en 1969. ¿Cuántos años hace que pasó eso? • Si estamos en 2016, lo que tienes que hacer es restar 1969 (que fue el año en que caminó en la Luna) a 2016: • $2016 - 1969 = 47$ • Respuesta: hace 47 años que Neil Armstrong caminó en la Luna
-------	---

- *application*, illustrates why the concept is useful and how it can be applied:

área	<ul style="list-style-type: none"> • tu escuela tiene dos patios de tamaño diferente y la directora quiere que la clase de deportes sea en el patio más grande. • Las medidas del primer patio son 25.3 m de largo por 17.45 m de ancho • Las medidas del segundo patio son 20.4 m de largo por 19.54 m de ancho • necesitas sacar el área de los dos patios para saber cuál es el más grande • área patio 1 = $25.3 \times 17.45 = 441.48 \text{ m}^2$ • área patio 2 = $20.4 \times 20.4 = 398.62 \text{ m}^2$ • si comparas las dos áreas observas que el patio 1 es más grande que el patio 2, entonces la respuesta es • R: la clase de deportes va a ser en el patio 1.
------	---

5 Results and evaluation

Although not large in size, COPSMAAT satisfies the needs for this project, as it contains all the available maths textbooks used in every primary school in Mexico. It therefore contains all maths terms that children find at least once in these years. In the next three subsections we summarise the results and provide an evaluation of the COPSMAAT from three perspectives: corpus terminology, terminology and knowledge transfer/education.

5.1 Corpus terminology

According to the classification criteria proposed by McEnery and Hardie (McEnery and Hardie 2012, 3-14) and Hunston (Hunston 2002, 14-6), COPSMAAT can be classified as follows: a) it is a corpus of written texts; b) it is domain-specific; c) it is exhaustive, as there are no other official primary school maths textbooks; d) it is representative of the language intended for children of a well-defined age bracket (6 - 12 year-olds); e) data sample reflects a snapshot of current maths curricula for primary school in Mexico; f) it is a corpus annotated with two main levels of annotation, each one of which has several layers of annotation.

The first level of annotation, as described above, inherited from COCIEM the original three layers of annotation corresponding to a) metadata: bibliographic reference, area to which the textbook belongs and school year in which the textbook is used; b) lexical: for each word, its lemma and POS; c) terminological: for each identified term, its lemma and POS.

A new layer was added to this level corresponding to additional terminological analyses exclusively applied to validated terms. These include: a) preferred term, in case there are two or more synonyms; b) abbreviation \rightarrow full form; c) symbol \rightarrow full form; d) sign \rightarrow full form; e) synonym \rightarrow preferred term; f) homonym (subdomain 1) \rightarrow homonym (subdomain 2); g) antonym 1 \rightarrow antonym 2.

The four layers within the first level of annotation are encoded in the corpus file itself.

In order to speed up processing times and to optimise the use of computer storage, the second level together with its two layers of annotation was stored in a separate file and applies only to validated terms. These are: 1) semantic and conceptual information; 2) practical information. The first layer includes: a) maths subdomain in which the term

in that context is used; b) definition or explanation, as considered appropriate; c) encyclopaedic information that is useful for a better understanding of the concept; d) example of the concept and, in the case of synonyms, different examples are provided for each one; e) related terms, where thematically related terms are cross referenced; f) illustration of the concept both in an abstract form and in the form of one or more of its applications.

The final layer of annotation includes practical information designed to provide children with directed activities to put the concept learned into practice. These include: a) exercise, explained step by step; b) how the concept is applied; c) how to obtain the value / result of an operation; d) problem, solved step by step; e) problem to be solved by the children.

In terms of corpus terminology, COPS MAT contains the information necessary to be an effective tool for mathematical knowledge transfer to primary school children.

5.2 Terminology

From the terminological perspective, COPS MAT was designed with the specific purpose to help children to understand basic mathematical concepts based on the official maths textbooks and on the curricula for the six years of primary school education. It therefore has a peculiar structure which differs from the traditional terminology corpus in several ways:

- It is not fully concept-oriented in the sense that each term is stored separately regardless of whether it is a preferred term, an accepted synonym, abbreviation, symbol or sign. However, all different designations for a concept are explicitly indicated within its internal structure.
- It has been designed to store additional information that in principle has nothing to do with terminology.
- Definitions have been drafted in a non-conventional form in order to make it easier for children to understand the concept in question by means of a set of short bulleted easy-to-follow sentences, each one explaining a characteristic of the concept. When convenient, additional features were added to the definition, such as “the symbol for x is y ” or “ x is necessary to calculate z ”.

In addition, from the terminological perspective, COPS MAT has made it possible to detect concepts that share characteristics; formal and concept variation within an individual textbook and across all nine textbooks; the explicit occurrence or absence of coordinate concepts; the inconsistent use of terms, etc.

5.3 Education/knowledge transfer

From the educational perspective, COPS MAT provides an excellent tool to evaluate textbooks and curricula through the use of language, what has been included and what should have been included but was left out. It is also possible to evaluate the sequence in which concepts are introduced to children, how many times a term is used in a school

year and in the following years to assess the relative importance given to that concept, and even typographical or other type of errors can be easily detected.

The fact that COPS MAT was supplemented with ad-hoc conceptual and practical information makes it unique and well suited for the purpose it was designed. In this way COPS MAT contains all the information needed to prepare and publish a terminographic product to complement textbooks with conceptual and practical information needed for the appropriate acquisition of mathematical concepts.

6 Further research and concluding remarks

As it can be deduced from this description, COPS MAT contains the necessary information for the *Diccionario de matemáticas para primaria*, that will be published later this year. Hopefully, it will help children to understand and acquire all those very basic mathematical concepts.

COPS MAT has opened a number of research possibilities for the near and mid-term future. Together with the corresponding science vocabulary learned at primary school constitutes the fundamental scientific vocabulary. It has particular quantitative and qualitative characteristics, that will be the topic of new research into the children's science and maths vocabularies.

Acknowledgement

I wish to thank Consejo Nacional de Ciencia y Tecnología (CONACYT) for their generous funding of this research project: *El vocabulario básico científico de México. Un estudio de sus características, componentes y difusión* (Clave 000000000220528).

References

1. Cabré, María Teresa. 1993. La Terminología. Teoría, metodología, aplicaciones. Barcelona: Atlántida/Empúries.
2. Cabrera-Diego, L.A., Sierra, G., Vivaldi, J., Pozzi, M. 2011. "Using Wikipedia to Validate Term Candidates for the Mexican Basic Scientific Vocabulary." In First International Conference on Terminology, Languages, and Content Resources (LaRC 2011). Seoul. 76-85.
3. Hunston, Susan. 2002. Corpora in Applied Linguistics. Cambridge: Cambridge University Press.
4. ISO 704:2009. Terminology work – Principles and methods. Geneva: ISO.
5. Lara, Luis Fernando, Ham Chande, Roberto, García Hidalgo, Isabel. 1979. Investigaciones Lingüísticas en Lexicografía. México D.F.: El Colegio de México.
6. McEnery, Tony, Hardie, Andrew. 2012 Corpus Linguistics. Cambridge: Cambridge University Press.
7. OECD. 2008. "Education at a Glance 2008". In OECD Briefing Note for Mexico. Accessed February 6, 2016. <http://www.oecd.org/edu/skills-beyond-school/41277868.pdf>
8. OECD. 2009. PISA 2009 Results: Executive Summary. Accessed February 6, 2016. <http://www.oecd.org/pisa/46643496.pdf>.

9. OECD. 2013. PISA 2012 Results: Executive Summary. Accessed February 6, 2016. <http://www.pisaresults-country-best-reading-maths-science/2013.pdf>.
10. Piaget, Jean. 1967. *Logique et Connaissance Scientifique*, Encyclopédie de la Pléiade. Paris: Éditions Gallimard.
11. Scott, M. 1996. *WordSmith Tools*, Oxford University Press, Oxford.
12. Vivaldi, Jorge, Cabrera-Diego, Luis Adrián, Sierra, Gerardo, Pozzi, María. (2012). "Using Wikipedia to Validate the Terminology found in a Corpus of Basic Textbooks". In *LREC 2012*. 820-27.
13. Vygotski, Len S. 1978. *Mind in Society. The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Enabling Linked-Data-Based Semantic Annotations – the Ontological Modeling of Semantics in the OntoLingAnnot Model

Antonio Pareja-Lora

Facultad de Informática, Universidad Complutense de Madrid, Spain
aplora@ucm.es

Abstract. This paper presents the ontology-based, linked-data-aware modeling of the vocabulary of Semantics included in the OntoLingAnnot model, aiming at a linguistic linked data compatible [semantic] annotation of texts. In particular, it introduces the different semantic units and attributes that can be used to annotate texts at the semantic level using the framework. These semantic units and attributes are included in the set of ontologies associated to OntoLingAnnot, whose main design assumptions are also described here. These main assumptions and the modeling performed has already helped different semantic (or linguistic, in general) annotations interoperate.

Keywords: ontology, linked data, linguistic, semantic, annotation, model, framework, unit, attribute, value, OntoLingAnnot.

1 Introduction

As discussed in Pareja-Lora (2012a, 2014), the traditional and most usual criterion to dissect Linguistics and linguistic annotation for their study and development is based on the concept of level, which divides Linguistics into, for example, Morphology, Syntax, Semantics, Discourse or Pragmatics. This division of Linguistics and its applications “has given rise to several good separate models of its different levels, which, nonetheless (and unfortunately), cannot interoperate and do not benefit from the advances of the others in most of the cases” (Pareja-Lora, 2014).

As shown also in Pareja-Lora (2014), even the standards carried out by the ISO TC37/SC4 subcommittee on linguistic annotation suffer somehow from this type of bias. For example, ISO 24611:2012 (Morpho-syntactic annotation framework – MAF), ISO 24615-1:2014 (Syntactic annotation framework, part 1 – SynAF), ISO 24613:2008 (Lexical markup framework – LMF) and the family of standards that try to cover the rest of semantic annotations (Semantic annotation framework – SemAF)¹

¹ That is, ISO 24617-1:2012 (SemAF-Time, ISO-TimeML- for the annotation of events and temporal expressions), ISO 24617-2:2012 (Dialogue acts, SemAF-DActs), ISO 24617-4:2014 (SemAF-SR – for the annotation of semantic roles), ISO/TS 24617-5:2014 (SemAF-

focus each on a separated and specific XML-based standard scheme for particular annotation levels or phenomena. Thus, these standards are not linked-data-aware and/or -ready and not necessarily fully interoperable.

Certainly, the level of coordination in the development of the ISO/SemAF family of standards is much higher than within the ISO/TC 37/SC 4 as a whole, resulting in a higher internal level of coherence and interoperability. In spite of this, the interoperability of the ISO/SemAF standards with the rest of ISO/TC 37 standards, in particular with those dealing with other aspects of semantics (such as lexical meaning² or the conceptual counterpart of terminological entries³), is in most cases insufficient. Besides, the number of semantic theories, approaches and/or phenomena standardized in the ISO/SemAF family is, unfortunately, not broad enough. For instance, (1) they do not cover the annotation of named entities as of yet; and (2) some of the standards in this family (e.g. SemAF-Time and ISOSpace) are mainly biased towards the annotation of English-specific syntactic and/or lexico-syntactic issues.

Therefore, a more comprehensive framework over linguistic annotation approaches in general and over semantic annotation approaches, theories and schemas in particular is still required. One of the approaches that more urgently need to be included in this comprehensive framework is, clearly, the linked data approach. In other words, this comprehensive approach should allow the generation and management of linked-data-based semantic annotations, as with the POWLA formalism (Chiarcos, 2012) or following the W3C's NIF 1.0 guidelines for linked data corpus creation (Brümmer, Ackermann & Dojchinovski, 2015).

This paper shows an example of such a comprehensive framework for linguistic and semantic annotation. More specifically, it shows the formalization of the semantic (annotation) level integrated into the OntoLingAnnot annotation framework; and it also presents the different semantic categories that can be used to annotate texts using the framework. These semantic categories are included in the ontologies (Gruber, 1993; Borst, 1997) associated to OntoLingAnnot.

The paper has been organized as follows. Section 2 states the background and the main assumptions underlying the OntoLingAnnot annotation framework. Then, the main units of its semantic level and the taxonomical relations holding between them are presented in Section 3. The attributes characterizing them are included in Section 4 (unfortunately, the corresponding values cannot be introduced here for the sake of space). Section 5 shows the evaluation results and the main contributions of the framework achieved so far; and finally, section 5 discusses the conclusions and the expected further work associated to this research.

DS – Discourse structure), ISO 24617-6:2016 (SemAF Principles), and ISO 24617-7:2014 (ISOSpace – for the annotation of spatial entities).

² Standardized in ISO 24613:2008 (LMF).

³ Standardized, for example, in ISO 16642:2003 – Terminological markup framework (TMF).

2 The OntoLingAnnot Model

The OntoLingAnnot model (as well as its ancestor, OntoTag – see Pareja-Lora (2012a; 2016)) was devised as a conceptual umbrella over several different linguistic theories, as well as a number of levels and approaches to linguistic annotation. OntoLingAnnot aims at providing an ontology-based, standardized, joint, structured, modular and interoperable framework for the annotation of morphological, syntactic, semantic, discourse-related and pragmatic phenomena (Pareja-Lora & Aguado de Cea, 2010; Pareja-Lora, 2012b, 2014).

Indeed, OntoLingAnnot was developed “following a comprehensive approach, which considers all these linguistic theories, levels and approaches to annotation altogether, not separately” (Pareja-Lora, 2014). Nevertheless, neat frontiers between the scopes of the different levels formalized had to be defined as well, “in order to (i) avoid redundancy; and (ii) identify clearly the interfaces between these levels” (Pareja-Lora, 2014).

This constitutes the backbone of the OntoLingAnnot annotation framework. In addition, the following assumptions were made when it was developed (Pareja-Lora, 2012b, 2014):

1. For the sake of annotation interoperability, and following ISO 24612:2012 (Linguistic annotation framework – LAF) and ISO 12620:2009 (Specification of data categories and management of a Data Category Registry for language resources – DCR), a clear differentiation had to be established between the linguistic data categories (LDCs) used to annotate and the format (or the way) in which these annotations are expressed.
2. For the same reason, LDCs had to be formalized as ontological items. This is the origin of OntoLingAnnot’s (as well as OntoTag’s) linguistic ontologies, which enable identifying and referring to each LDC by means of its own Uniform Resource Identifier (URI) (one of the requirements of ISO 24619:2011 – Persistent identification and sustainable access, PISA).
3. Whenever possible, these ontological items formalizing LDCs, should be linked to the LDCs included in ISocat (the implementation of ISO 12620:2009 – DCR)⁴.
4. To facilitate an RDF-based representation of annotations, and also to avoid redundancy and facilitate modularization (Pareja-Lora, 2012a, 2012b), each LDC had to be classified as a linguistic unit (such as “noun”), as a linguistic attribute (such as “gender”), as a linguistic value (such as “neuter”), or as a linguistic relation (such as “syntactic dependency” or “syntactic constituency”, cf. Pareja-Lora (2012c)).
5. Basically, annotating a text entails attaching to it a set of annotation triples <Linguistic Unit, Linguistic Attribute, Linguistic Value>. These annotation triples can then be implemented as RDF triples <Subject, Predicate, Object>, in which the corresponding linguistic units (i.e., subjects), attributes (i.e., predicates), and values (i.e., objects), are conveniently formalized as classes or individuals of one or more ontologies (also this assumption is ISO 24612:2012-compliant).

⁴ <http://www.isocat.org/>

6. The framework had to (i) maximize the coverage as for the phenomena that it contemplated and, thus, of the LDCs that it included; and (ii) be flexible and scalable enough to allow its users to select the set of categories included in their annotations (this is enabled in the model by means of the implementation of LDCs using one or more ontologies).
7. The previous assumption required (a) following an eclectic and/or non-theory-biased approach for the selection of the LDCs that had to be finally included in the framework; (b) defining a coherent and theory-neutral terminology for the designation of the categories; (c) accompanying each LDC with as many synonyms and/or labels as needed (that is, when several theories referred to the same phenomenon in a different way); and (d) adding many new concepts or labels as needed, in order to link together the terms coming from different and complementary theories or approaches, but referring to the same linguistic item.

This completes the specification of the main pillars and assumptions underlying the OntoLingAnnot framework. The following two sections present the main different classes and individuals that have been included in the semantic modules of OntoLingAnnot's ontologies. They represent the main categories that can be used for the semantic annotation of texts within this framework. They are presented in two dedicated sections, according to their linguistic type, namely semantic units and semantic attributes.

3 The Main Semantic Units in OntoLingAnnot

The main semantic units modeled in OntoLingAnnot are shown in Table 1.

Table 1. The top-level semantic units in OntoLingAnnot

TOP-LEVEL CONCEPTS	SEMANTIC CONCEPTS	
Semantic Unit	Lexical Meaning Unit (Syntactic-Semantic Interface Unit, LMU)	Simple Lexical Meaning Unit (Morphosyntactic-Semantic Interface Unit, SiLMU)
		Complex Lexical Meaning Unit (Phrasal Meaning Unit, Phraseme, CoLMU)
	Sub-Lexical Meaning Unit (Sememe, Morphological-Semantic Interface Unit, SubLMU)	
	Supra-Lexical Meaning Unit (Discourse Propositional Unit, Proposition, SupraLMU)	

As shown in this table, a Semantic Unit can be further subclassified as a Sub-Lexical Meaning Unit, a Lexical Meaning Unit or a Supra-

Lexical Meaning Unit⁵. In order for the definitions of some of these units to be more easily understood, two other definitions need to be introduced beforehand. On the one hand, a *simple (mental) construct* is assumed to be a single (mental) concept or idea in OntoLingAnnot; on the other hand, a *complex (mental) construct* is assumed to be a (mental) construct that involves two or more interrelated concepts or ideas.

Thus, firstly, the class **Lexical Meaning Unit** can be thought of as a type of Semantic Unit whose syntactic realization lacks a Clause or Sentence rank, and which formalizes either (1) a simple mental construct that can be realized by means of one or more morphosyntactic units; or else (2) a complex (mental) construct that is usually realized by means of a single lexical (but plurilexematic) unit, that is, a fixed, lexicalized and/or decompositional (*i.e.*, non-compositional) combination of morphosyntactic units.

Secondly, the class **Sub-Lexical Meaning Unit** is a Semantic Unit that can be lexicalized by a certain kind of morph or, in other words, whose morphosyntactic projection is a morph. For example, the English Word (Form) ‘trees’ has two morphs: the Root, ‘tree’, and the pluralizing Affix ‘-s’, which carries the meaning [+ plural] (*cf.* Crystal, 1992). Hence, the English Word (Form) ‘trees’ has two associated sub-lexical meaning units, each one corresponding to one of its constituent morphs. This subclass might seem redundant with respect to Lexical Meaning Unit, taking into account that, at least in most Western languages, almost any Affix can be re-expressed in terms of a Stem, or realizes a (mental) construct. However, this subclass was included for completeness sake, with respect to (1) the terminology found in the Semantics literature, (2) the possible phenomena that might exist in those languages unknown by the author, and (3) the subclassification criterion of Lexical Meaning Unit itself.

Thirdly, the class **Supra-Lexical Meaning Unit** (Propositional Unit) can be described as a type of Semantic Unit (1) that formalizes a complex (mental) construct (2) whose syntactic projection is a Clause or a (Simple) Sentence, and (3) whose meaning can be calculated compositionally from the meanings of its components. Therefore, a Propositional Unit can also be regarded of as an aggregation or the composition of some interrelated semantic components that, altogether, constitute a higher-level Semantic Unit, which results from the straightforward composition of the meanings of its components.

Table 1 also shows the subclasses of Lexical Meaning Unit, namely, Simple Lexical Meaning Unit, and Complex Lexical Meaning Unit.

Concerning the first subclass of Lexical Meaning Unit, a **Simple Lexical Meaning Unit** is a type of Semantic Unit that formalizes a single (mental) construct, *i.e.*, a single concept or idea. For example, the Simple Lexical Meaning Unit associated to the Spanish word ‘árbol’ would be the class repre-

⁵ In such a subclassification table, (i) representing a concept, C1, in a column on the right of another concept (or class), C2, and on its same row, means that C1 is a subconcept or a subclass of C2; and (ii) the terms represented between parentheses in some cells are synonyms, acronyms or actual designators of the concept in a given linguistic theory or approach.

senting the concept Tree in a given ontology. Also the phrase ‘The President of the USA by 2007’ itself refers to a single idea, the mental representation of a particular Entity of the real world, Mr. George Walker Bush. Therefore, a single (mental) construct (that is, a Simple Lexical Meaning Unit) can be realized by just one word (such as ‘Bush’) or by a syntactic combination of words (such as ‘The President of the USA by 2007’).

Yet, as hinted above, there are other examples of lexical units whose meaning cannot be reduced to a single (mental) construct or idea. This type of lexical units has been formalized by means of the class **Complex Lexical Meaning Unit**. It encompasses all those fixed, lexicalized and/or decompositional (*i.e.*, non-compositional) combinations of words that express a (mental) construct that cannot be reduced to a single idea. Complex lexical meaning units, therefore, consist of two or more simple lexical meaning units, together with the relationships established between them in order to build a higher-order (mental) construct. For example, the Spanish expression ‘*dar un golpe de Estado*’, and its corresponding translations into English (‘*stage a coup d’état*’), French (‘*donner un coup d’état*’), Italian (‘*dare un colpo di stato*’) and German (‘*einen Staatsstreich / Putsch inszenieren*’) can be considered each an *Instance-Of* Complex Lexical Meaning Unit, since they involve an Action (which might be referred to as *Perform*) in which the semantic Object is fixed (the ontological class(es) formalizing the concept(s) associated to *coup d’état*). In some languages, a **Complex Lexical Meaning Unit** can also be expressed by a single word: for example, the Spanish word ‘hachazo’ (≡ ‘blow of/with an axe’ in English; ‘coup d’hache’ in French; and ‘colpo di scure’ in Italian), from a cross-linguistic and lexical point of view, cannot be reduced to a single mental construct. Hence, it should be annotated as a Complex Lexical Meaning Unit as well.

With respect to the boundaries between the classes Complex Lexical Meaning Unit and Proposition, complex lexical meaning units cannot be considered propositions *per se*. On the one hand, they are incomplete propositions (Corpas-Pastor, 1996), that is, they lack some of the elements that characterize a Proposition. On the other hand, they are fixed and lexicalized to some extent and, hence, they are not as variable and compositional as true propositions.

All these semantic units have been conveniently subclassified in OntoLingAnnot’s ontologies, but their full subclassification cannot be presented here for the sake of space. Nevertheless, the main subclasses of Simple Lexical Meaning Unit are included in Also the subclassification of Propositional Component is introduced here, in order for the semantic attributes and values presented in Section 0 to be better understood. This subclassification is shown in Table 3. The main concepts in this table (Entity, Action, and Quality) have been extracted from Lyons (1977). For this author, these main concepts represent the ontological basis of Semantics and Grammar.

Table 2. This table also shows where the concept Synset⁶ has been placed in OntoLingAnnot’s ontologies.

Also the subclassification of Propositional Component is introduced here, in order for the semantic attributes and values presented in Section 0 to be better understood. This subclassification is shown in Table 3. The main concepts in this table (Entity, Action, and Quality) have been extracted from Lyons (1977). For this author, these main concepts represent the ontological basis of Semantics and Grammar.

Table 2. The subclasses of Simple Lexical Meaning Unit in OntoLingAnnot

Simple Lexical Meaning Unit	Sense	Synset
	Propositional Component	
	Other Simple Lexical Meaning Unit	

Table 3. The subclasses of Propositional Component in OntoLingAnnot

Propositional Component	Entity ⁷	Named Entity		
		Generic Entity	Concrete Entity	{Location, Material, Artifact, Food, Physical Object, Organic Object, Living Entity, Substance}
			Abstract Entity	{Domain, Time, Moral Standard, Cognitive Fact, Movement Of Thought, Institution, Convention}
		Action (Process, State Of Affairs, SoA)		
	Quality	Property (Entity Quality)		
		Circumstance (Process Quality)		

Firstly, Entity has been further subclassified into Named Entity and Generic Entity in order to allow for a suitable interoperability of named entity annotations and other semantic annotations (Aguado de Cea, Álvarez de Mon y Rego & Pareja-Lora, 2009). Named Entity has been subclassified in OntoLingAnnot according to the MUC-7 (Chinchor, 1997) and ACE (Doddington *et al.*, 2004) initiatives; and Generic Entity has been subclassified according to the SIMPLE project (see the details in Pareja-Lora, 2012a).

Secondly, some *synonyms* for Action (State Of Affairs and Process) have been linked to this concept in OntoLingAnnot. These *synonyms* are the terms used in other linguistic grammars and theories to refer to the same concept. On the

⁶ This concept is crucial in both WordNet and EuroWordNet, which are the *de facto* standard tagsets for the semantic annotation of senses nowadays.

⁷ This concept is referred to as Participant in Halliday (1994; 1996).

one hand, the term *State Of Affairs* (as well as its corresponding attributes and values) has been extracted from Dik (1989). On the other hand, the term *Process* (as well as its subclassification, not presented here for brevity) has been derived from Halliday (1994; 1996).

Thirdly, the subclassification of *Quality* into *Property* and *Circumstance* and their subclasses has been formalized according to the SIMPLE project (see the details in Pareja-Lora, 2012a), like *Generic Entity*.

4 The Main Semantic Attributes in OntoLingAnnot

This section presents the concepts of the OntoLingAnnot ontologies that formalize the linguistic attributes of semantic units (that is, semantic attributes). The main subclasses of *Semantic Attribute* in OntoLingAnnot are *Lexical Function*, *Semantic Feature*, *Lexical Meaning Unit Attribute*, *Propositional Component Attribute*, *State Of Affairs Attribute* (*Action Attribute*), *Quality Attribute*, and *Other Semantic Attribute*. The individuals already identified for them are shown in Table 4, together with the sources where they have been extracted. They cannot be further discussed here for the sake of space.

Table 4. The individuals of semantic attributes in the LAO

SEMANTIC ATTRIBUTE CONCEPTS	SEMANTIC ATTRIBUTE INDIVIDUALS
Lexical Function	Me••uk’s Lexical Function
Semantic Feature (Aarts & Calbert, 1979)	{isConcrete, isLiving, isHuman, isMale, isAnimal, hasShape, isArtifact, isPerceptible, isState, isPhysical, isDimensional, isVolitive ⁸ , isAttribute, isEvaluative}
Lexical Meaning Unit Attribute (Cruse, 1986; 1997; Corpas-Pastor, 1996; SIL, 2016)	{isSemanticallyTransparent, isIdiomatic, isSubstitutable, isModifiable, isLexicallyFixed} ⁹
Propositional Component Attribute	{isInstanced, hasParticipantType, hasSemanticRole}

⁸ This instance was referred to as “ACTION” in the original source, *i.e.*, Aarts & Calbert (1979). However, the original term leads to a misunderstanding of its meaning, which can be better expressed by means of the term chosen for its formalization in the present ontology.

⁹ On the one hand, the attribute *isSemanticallyTransparent* is the antonym of the attribute *isIdiomatic*; on the other hand, the attribute *isLexicallyFixed* is equivalent to the negation of both *isSubstitutable* and *isModifiable*. Therefore, the inclusion of all of them in a given annotation schema is not recommended, but optional.

(Halliday, 1994; 1996; Gildea & Jurafsky, 2002)	
State Of Affairs Attribute (Dik, 1989)	{isDynamic, isTelic, isMomentaneous, hasController, isExperiential}
Quality Attribute (Lázaro-Carreter & Tusón, 1978)	{isAttributive, isPredicative}

5 Evaluation and Contributions of the Present Work

To the best of our knowledge, this is the first and only attempt to formalize in an ontology the vocabulary and/or the terminology associated to Semantics thus far. On the one hand, neither the OLiA ontologies¹⁰ nor the GOLD ontology¹¹ (the main ontologies formalizing linguistic terminology and/or knowledge) have a particular module for the representation of semantic knowledge. On the other hand, the other main LDC repository broadly known and used within the linguistic linked data community (i.e., ISOcat) only includes a few LDCs dealing with Semantics, chiefly related to ISO 24613:2008 (LMF). Therefore, the **ontological (and the linguistic linked-data cloud) gap-filling** of the semantic level of OntoLingAnnot's ontologies is irrefutable. Besides, the claims in Pareja-Lora (2014) about the pragmatic level of OntoLingAnnot, namely, (i) scalability; (ii) extensibility; (iii) interoperability; (iv) standard-compliance; and (v) usability; can also be made here about its semantic level, since they derive mainly from the formal properties of the OntoLingAnnot framework.

Another interesting fact about the OntoLingAnnot ontologies is that they are currently undergoing a process of review and restructuring, in order to adapt them to some interesting functionalities and mechanisms of OWL 2.0 that were not present in OWL 1.0 (one of the main languages originally used to implement these ontologies). For example, developing property hierarchies was not feasible in OWL 1.0, the only version of OWL that existed when these ontologies were created. This undoubtedly affected the way they were conceived and structured. For instance, implementing linguistic attribute hierarchies required transforming linguistic attributes into classes beforehand. That is, representing these data categories required changing their ontological status. Since this restriction is not applicable anymore, a new version of the ontologies that avoids this problem (amongst others) is being developed. Some additional experiments are being carried out as well in order to (a) find out if all other similar development restrictions are still valid and (b) provide a more language-independent (and, thus, more conceptually appropriate) version on the ontologies in the future. As soon as this version has been conveniently evaluated, it will be made available and accessible at a public URI (to be determined – most possibly, within GitHub¹²).

¹⁰ <http://acoli.cs.uni-frankfurt.de/resources/olia/>

¹¹ <http://linguistics-ontology.org/gold/2010>

¹² <https://github.com/>

6 Conclusions & Further Work

This paper has introduced the semantic annotation level of the OntoLingAnnot (linguistic) annotation framework, focusing on its semantic categories, which are included as concepts and individuals of its ontologies. The corresponding ontological modules of OntoLingAnnot formalize the different semantic units, attributes and values (as well as relationships) identified in the literature so far, and constitute a coherent distribution and structuring of these semantic categories as for their use in (semantic) annotation.

As shown in the previous sections, this is the first and only ontological and/or linked-data-aware conceptualization of Semantics thus far and, hence, it is an important contribution per se to the areas of Terminology, Knowledge Engineering, Ontological Engineering and Linguistic Annotation, as well as to the linguistic linked data cloud. Besides, no other model or framework accounts globally and coherently for such a number of semantic phenomena and categories as those formalized and included in OntoLingAnnot's ontologies, which is another important contribution to the areas aforementioned. In addition, as mentioned in the previous section, this approach is also scalable, extensible, interoperable, standard compliant and highly (re)usable.

However, the mapping of OntoLingAnnot's semantic categories into/onto the DCR is still pending. This remains as further work to be accomplished shortly.

References

- Aguado de Cea, Guadalupe, Inmaculada Álvarez de Mon y Rego, and Antonio Pareja-Lora. 2009. Una visión interdisciplinar de la anotación semántica. In *Terminología y sociedad del conocimiento*. Berlin: Peter Lang, pp. 219-254.
- Aarts, Jan M. G., and Joseph P. Calbert. 1979. *Metaphor and Non-Metaphor: the Semantics of Adjective Noun Combinations*. Tübingen: Max Niemeyer Verlag.
- Austin, John L. 1975. *How to do things with words*. Oxford: Oxford University Press.
- Borst, Willem N. 1997. *Construction of Engineering Ontologies*. PhD thesis. University of Twente. Enschede. Netherlands.
- Brümmer, Martin, Markus Ackermann, and Milan Dojchinovski. 2015. *Guidelines for Linked Data corpus creation using NIF (v. 1.0)*. W3C Community Group Draft Report. 29 September 2015 [Available online at <http://bpmlod.github.io/report/nif-corpus/index.html>, accessed on 2016/04/06].
- Chiarcos, Christian. 2012. Interoperability of Corpora and Annotations. In Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), *Linked Data in Linguistics. Representing Language Data and Metadata*. Heidelberg: Springer, pp. 161-179.
- Chinchor, Nancy. 1997. MUC-7 Named Entity Task Definition, Version 3.5 [Available on line at: http://www-nlpir.nist.gov/related_projects/muc/proceedings/ne_task.html, accessed on 2016/04/06].
- Corpas-Pastor, Gloria. 1996. *Manual de fraseología española*. Madrid: Gredos.
- Cruse, D. Alan. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.
- Crystal, David. (1992). *A Dictionary of linguistics and phonetics* (3rd Edition). Oxford: Blackwell Publishers.

- Dik, Simon C. 1989. *The Theory of Functional Grammar: The structure of the clause*. Dordrecht: Foris Publications.
- Doddington, G., A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The Automatic Content Extraction (ACE) Program. Tasks, Data, and Evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'2004)*, Lisboa.
- Gildea, D., and D. Jurafsky. 2002. Automatic Labelling of Semantic Roles. *Computational Linguistics*, 28(3): 245–288.
- Halliday, Michael A.K. 1994 (1996). *An Introduction to Functional Grammar (2nd Edition)*. London: Arnold.
- ISO 12620:2009. *Terminology and other language and content resources – Specification of data categories and management of a Data Category Registry for language resources*. International Organization for Standardization (ISO).
- ISO 16642:2003. *Computer applications in terminology – Terminological markup framework*. International Organization for Standardization (ISO).
- ISO 24611:2012. *Language resource management – Morpho-syntactic annotation framework (MAF)*. International Organization for Standardization (ISO). International Organization for Standardization (ISO). International Organization for Standardization (ISO).
- ISO 24612:2012. *Language resource management – Linguistic annotation framework (LAF)*. International Organization for Standardization (ISO).
- ISO 24613:2008. *Language resource management. Lexical markup framework (LMF)*. International Organization for Standardization (ISO).
- ISO 24615-1:2014. *Language resource management – Syntactic annotation framework (SynAF) – Part 1: Syntactic model*. International Organization for Standardization (ISO).
- ISO 24617-1:2012. *Language resource management – Semantic annotation framework (SemAF) – Part 1: Time and events (SemAF-Time, ISO-TimeML)*. International Organization for Standardization (ISO).
- ISO 24617-2:2012. *Language resource management – Semantic annotation framework (SemAF) – Part 2: Dialogue acts*. International Organization for Standardization (ISO).
- ISO 24617-4:2014. *Language resource management – Semantic annotation framework (SemAF) – Part 4: Semantic roles (SemAF-SR)*. International Organization for Standardization (ISO).
- ISO/TS 24617-5:2014. *Language resource management – Semantic annotation framework (SemAF) – Part 5: Discourse structure (SemAF-DS)*. International Organization for Standardization (ISO).
- ISO 24617-6:2016. *Language resource management – Semantic annotation framework – Part 6: Principles of semantic annotation (SemAF Principles)*. International Organization for Standardization (ISO).
- ISO 24617-7:2014. *Language resource management – Semantic annotation framework – Part 7: Spatial information (ISospace)*. International Organization for Standardization (ISO).
- ISO 24619:2011. *Language resource management – Persistent identification and sustainable access (PISA)*. International Organization for Standardization (ISO).
- Gruber, Thomas R. 1993. A Translation Approach to Portable Ontologies. In *Journal on Knowledge Acquisition*, Vol. 5(2): 199–220.
- Lázaro Carreter, Fernando, and Vicente Tusón Valls. (1978). *Curso de lengua española: memorándum para el profesor*. Madrid: Anaya.
- Lyons, John. 1977. *Semantics*. Cambridge: Cambridge University Press.
- Pareja-Lora, Antonio and Guadalupe Aguado de Cea. 2010. Modelling discourse-related terminology in OntoLingAnnot's ontologies. In Úna Bhreathnach and Fionnuala Barra-Cusack

- (eds.), *Presenting terminology and knowledge engineering resources online: models and challenges (TKE 2010)*. Dublin, July 2010, pp. 547 - 574. Dublin: Ass. for Terminology and Knowledge Transfer (Gesellschaft für Terminologie und Wissenstransfer, GTW).
- Pareja-Lora, Antonio. 2012a. *Providing Linked Linguistic and Semantic Web Annotations – The OntoTag Hybrid Annotation Model*. Saarbrücken: LAP – LAMBERT Academic Publishing.
- Pareja-Lora, Antonio. 2012b. OntoLingAnnot's Ontologies: Facilitating Interoperable Linguistic Annotations (Up to the Pragmatic Level). In Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), *Linked Data in Linguistics. Representing Language Data and Metadata*. Heidelberg: Springer, pp. 117-127.
- Pareja-Lora, Antonio. 2012c. OntoLingAnnot's LRO: An Ontology of Linguistic Relations. In *Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012)*. Madrid, June 2012, pp. 49-64. Madrid: Universidad Politécnica de Madrid [Available online at http://oeg-lia3.dia.fi.upm.es/c/document_library/get_file?uuid=2c09de2b-51f1-491e-a244-444e725582f9&groupId=10157, accessed on 2016/05/20].
- Pareja-Lora, Antonio. 2014. The pragmatic level of OntoLingAnnot's ontologies and their use in pragmatic annotation for language teaching. In J. Arús, M.E., Bárcena, and T. Read (eds.) *Languages for Special Purposes in the Digital Era. Series: Educational Linguistics, Vol. 19*, pp. 323-344. Switzerland: Springer International Publishing Switzerland.
- Pareja-Lora, Antonio. 2016. "Enabling automatic, technology-enhanced assessment in language e-learning – Using ontologies and linguistic annotation merge to improve accuracy". In Elena Martín-Monje, Izaskun Elorza, Blanca García Riaza (eds.) *Technology-Enhanced Language Learning for Specialized Domains – Practical Applications and Mobility*, pp. 102-126. Oxon & New York: Routledge.
- SIL. 2016. Glossary of Linguistic Terms (Eugene E. Loos, Susan Anderson, Dwight H. Day (Jr.), Paul C. Jordan and J. Douglas Wingate, eds.). [Available online at <http://www-01.sil.org/linguistics/GlossaryOfLinguisticTerms/>, accessed on 2016/04/06].

Terminoteca RDF: a Gathering Point for Multilingual Terminologies in Spain

Julia Bosque-Gil, Elena Montiel-Ponsoda, Jorge Gracia, and Guadalupe Aguado-de-Cea

Ontology Engineering Group, Universidad Politécnica de Madrid, ETSI Informáticos,
Campus de Montegancedo sn, 28660 Boadilla del Monte, Madrid

Abstract. Terminological resources can greatly benefit from techniques that enable their transformation and linking to become a navigable graph of linked language resources, published on the Web according to the linked data paradigm. In this work we present Terminoteca RDF, a prototype that aims to lay the foundations of a repository of linked multilingual terminologies of official languages in Spain. In this contribution we describe the model adopted to represent such terminologies in linked data and spell out the tasks followed in the transformation process from proprietary formats to RDF, namely, data exploration, URI naming strategy definition, data modelling, and RDF generation and linking.

Keywords: terminological resources, linked data, lemon-ontolex, multilingualism

1 Introduction

Terminology work not only consists in identifying and defining the terms used in professional and scientific settings to create terminological resources, but also in taking advantage of representation formats, management systems or technologies that can impact the use of those resources and assist users. In the era of *linked open data* (Bizer et al., 2009), terminology can greatly benefit from the publishing of terminological resources in Semantic Web formats and, most importantly, from the linking to other terminological, linguistic or content resources that can significantly enrich the information they contain. For instance, the Terminology Coordination Unit of the European Parliament (TermCoord), as a good representative of terminology practices and work, is committed to “extend terminologies to new horizons” (Maslias, 2014), and not only recognizes the importance of adapting to new formats and technologies, but also promotes it from the recently inaugurated TermCoord platform.¹

In this context, linked data technologies (Bizer et al., 2009) constitute a major opportunity for representing, sharing, interlinking, and accessing terminological information. According to this paradigm, data has to be described according to the Resource Description Framework (RDF) data model (Manola and Miller,

¹ <http://termcoord.eu/>

2004). This allows computers to easily interpret data as resources uniquely identified at Web scale. Then, that data has to be linked to data in related resources. Finally, data can be retrieved and manipulated by using Web standards such as the SPARQL² query language.

Currently, a large number of language resources are being transformed to this new paradigm. This has enabled the emergence of the so called Linguistic Linked Open Data (LLOD) cloud.³ The LLOD cloud contains monolingual and multilingual language resources such as dictionaries, typological databases, thesauri and even corpora.

In this work we describe our contribution to the LLOD cloud with the creation of *Terminoteca RDF*, a collection of interlinked multilingual terminologies in Spain. Currently, *Terminoteca RDF* contains two sets of terminological resources: *Terminesp*⁴, a multilingual terminological database created by the Spanish Association for Terminology, AETER, by extracting the terminological data from the UNE documents produced by AENOR (Asociación Española de Normalización y Certificación) on the one hand, and a set of freely available terminological databases from the Catalan Terminological Centre, *TERMCAT*⁵, on the other. These terminological resources were developed independently and following non-standard formats. As result of their inclusion in *Terminoteca RDF*, their data are currently represented in RDF and are connected in a common graph and exposed as linked data on the Web. The data is accessible through an SPARQL endpoint and also via a web interface.⁶ In a first stage, we would like *Terminoteca RDF* to become a platform of reference for those languages that are official in Spain (Spanish, Catalan, Basque and Galician), but it already contains term descriptions in many other languages (including Latin).

The paper has been structured as follows. A state of the art on linked terminologies is presented in section 2. In section 3 we provide a description of the model we have adopted to represent the data in *Terminesp* and the *TERMCAT* databases, which is the *lemon-ontolex* model, a *de facto* standard for describing language resources in the Web of Data. Section 4 is devoted to the different tasks we have completed in the transformation and linking processes. Section 5 highlights the benefits of browsing and navigating linked terminologies and introduces future lines of work.

2 Related Work

The importance of integrating terminologies and other types of language resources is demonstrated not only by the exponential growth of the LLOD cloud, but also by the great amount of projects that aim at bringing together quality language resources, and provide a single access point. The TermCoord unit has

² <http://www.w3.org/TR/rdf-sparql-query/>

³ <http://linguistic-lod.org/llod-cloud>

⁴ <http://www.fundeu.es/tema/terminesp/>

⁵ <http://www.termcat.cat/>

⁶ <http://linguistic.linkeddata.es/terminoteca>

been involved in several projects in the last years. Recently, a workshop⁷ was hosted at the European Parliament in Luxembourg with the aim of mapping resources created by the European institutions, namely IATE⁸, the well-known InterActive Terminology for Europe, or EUROVOC⁹, a multilingual and multidisciplinary thesaurus, to external semantic resources, specifically, the multilingual knowledge base BabelNet (Navigli and Ponzetto, 2010). In a similar trend, TermCoord and TERMCAT have reached an agreement in order to enrich and update the contents of the internal version of IATE with relevant terminological data in Catalan¹⁰, just to mention some recent cases.

As for terminological resources in RDF linked to other resources, we refer to IATE RDF, a data export of the IATE term base¹¹ that is part of the LLOD cloud, and which has been linked to the European Migration Network (EMN) glossary (Cimiano et al., 2015). As mentioned in that paper, there is no standard format for publishing terminologies as RDF. Most terminologies follow the TBX (TermBased Exchange Format) model, an XML-based format to represent a set of terms grouped by language that designate a concept.¹² In fact, in this same work, the authors develop a converter from TBX to RDF based mainly on the *lemon-ontolex* model¹³, and a set of best practices for transforming terminologies in TBX into the linked data format.¹⁴

It is also worth mentioning the Simple Knowledge Organization System (SKOS) (Miles and Bechhofer, 2009) data model for representing the information in knowledge organization systems (thesauri, taxonomies, classification schemes and subject heading systems). In this sense, we find several well-known thesaurus in the LLOD cloud which have been exposed as linked data according to this data model, namely the SKOS version of the Food and Agriculture Thesaurus AGROVOC¹⁵ (currently linked to 16 other vocabularies and resources) or the GEMET¹⁶ thesaurus of the European Environment Agency.

For the sake of interoperability with other language resources published as linked data, and in accordance with Cimiano et al. (2015), we have also decided to comply to Semantic Web standards and reuse *lemon-ontolex* and SKOS to describe the terminological resources contained in Terminoteca RDF. By complying to *lemon-ontolex* we can also take advantage of the classes in that model that account for the relations between lexical entries or terms in the same or different languages, i.e., to represent term variants and translations. Even though the po-

⁷ <http://termcoord.eu/2016/03/when-linguistics-and-it-meet-babelnet-workshop-at-the-ep/>

⁸ <http://iate.europa.eu/>

⁹ <http://eurovoc.europa.eu/>

¹⁰ <http://termcoord.eu/2016/03/termcoord-and-termcat-in-close-cooperation-to-enrich-iate-content/>

¹¹ <http://tbx2rdf.lider-project.eu/data/iate>

¹² <http://www.tiu.ac.jp/org/openforum2006/slides/P5.pdf>

¹³ https://www.w3.org/community/bpmlod/wiki/Converting_TBX_to_RDF

¹⁴ <https://www.w3.org/2015/09/bpmlod-reports/multilingual-terminologies/>

¹⁵ <http://aims.fao.org/standards/agrovoc/linked-open-data>

¹⁶ <http://www.eionet.europa.eu/gemet>

tential of *lemon-ontolex* is not fully exploited in this work, we have adopted this model for several reasons. First, *lemon* subsumes most features of SKOS (e.g. preferred and alternative labels) and SKOS-XL (Miles and Bechhofer, 2009) reification of labels. Secondly and in contrast to other models, the *vartrans* module frames translations as relations between lexical senses and not between labels, which we deem more accurate linguistically. Lastly, sticking to *lemon-ontolex* will allow us to better integrate these resources with other lexicographic resources already available as *lemon* datasets (e.g. Apertium dictionaries) as well as to enrich them with information coming from external sources and pertaining to the relation between those lexical senses, reified in the *vartrans* module as *vartrans:SenseRelation* (with *vartrans:TerminologicalRelation* and *vartrans:Translation* as subclasses).

3 *lemon-ontolex*: a model for terminologies in RDF

In this section we briefly present the main features of the *lemon-ontolex* model¹⁷ and focus on one of the modules that constitute it, namely, the *vartrans* module, which serves to represent term variants and translations. *lemon-ontolex* is the resulting work of the W3C Ontology Lexica Community Group since 2011 to build a rich model in RDF that serves as interface between an ontology and the natural language descriptions that lexicalise the knowledge represented and structured in that ontology.

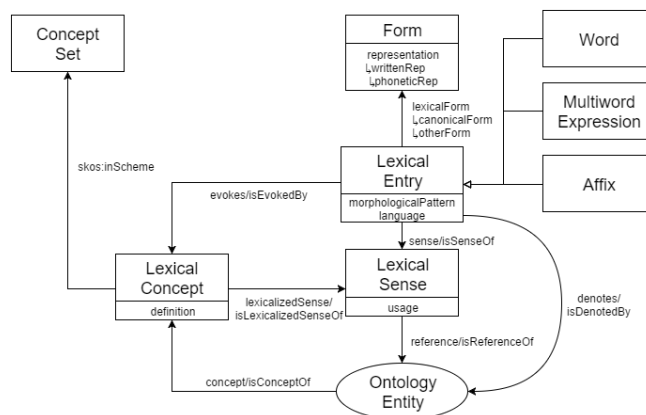


Fig. 1. The *lemon-ontolex* core

Figure 1 depicts the main classes and properties of the *lemon-ontolex* core. The main class of the core is the class **LexicalEntry**. Lexical entries can be linked to ontology entities in two ways: directly by the **denotes** property, or by

¹⁷ https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

means of an intermediate element called `LexicalSense`, which is intended to capture the particular sense of a word when referring to an ontology entity. The latter element allows us to attach the additional pragmatic properties describing under which conditions (context, register, domain, etc.) the lexical entry can be regarded as having the ontological entity as meaning. We can also represent the fact that a certain lexical entry evokes a mental concept or unit of thought that can be lexicalised by a given collection of senses. In that case, we would use the `LexicalConcept` class, which is a subclass of `skos:Concept`.

As for the *vartrans* module¹⁸, it has been developed with the aim of accounting for denominative variation and translations, although it also covers other types of lexico-semantic relations (such as synonymy, antonymy, or hyperonymy-hyponymy). Lexical relations are relations that can be established among lexical entries and/or forms concerning the surface form of a term and encoding morphological and orthographical variation, among other aspects. Terminological relations, as well as translations, are relations that can be established among senses within lexicons in the same or in different languages. Broadly speaking, we can say that term variants are pragmatically caused because of dialectal, chronological, discursive, dimensional, or formality reasons. The reasons that cause that variation are usually not captured in the ontology, but can be accounted for at the lexical sense level, and explicitly defined in the category property that describes the type of lexico-semantic relation in question.

4 Methodology

According to well established methodologies for publishing multilingual linked data on the Web (Vila-Suero et al., 2014), we followed these tasks to generate the linked data contained in Terminoteca RDF: source data exploration, URI naming strategy definition, data modelling and RDF generation and linking.

4.1 Source data exploration

Terminesp is a terminological database in Spanish whose terms were extracted from the UNE documents (similar to ISO Standards) from AENOR (Asociación Española de Normalización y Certificación), which, in turn, were produced by committees of experts in different domains. The database consists of more than thirty thousand terms for which definitions, norms, definition notes, and translations to Italian, English, German, French and Swedish are provided, along with scientific denominations in Latin. Some definitions and definition notes in these languages are included as well. The conversion of Terminesp to RDF has already been addressed in the literature (Gracia et al., 2014; Bosque-Gil et al., 2015) and it has been taken as basis for the work presented here.

TERMCAT repositories gather terminologies in Catalan divided by domain and are available through the portal *Terminologia Oberta*.¹⁹ Since the overlap

¹⁸ See ‘variation and translation’ section at https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

¹⁹ http://www.termcat.cat/en/Terminologia_Oberta/

of TERMCAT entries with those of Terminesp was crucial in order to show the potential of linking terminologies on the basis of the linked data paradigm, we selected domains that are shared by both terminologies, namely, *Internet*, *Telecommunications*, and *Electronics*. These TERMCAT domain lexica include translations from Catalan to Spanish, English, and, if available, to French, as well as synonyms, abbreviated forms, initialisms, and obsolete or dismissed denominations (the latter not modelled in the current work). Morphological information about part of speech and gender are also taken into consideration.

An example of a simple entry in a TERMCAT lexicon is presented below (Example 1.1). The term *administrador -a de xarxes* ‘network administrator’ (masc. and fem.) is described in the domain lexicon *Internet i Societat de la Informació*, record 32. The part of speech of both the main term and the synonym (*remissio* ‘remission’ of type *terme pral.*) are provided in the attribute *categoría* ‘category’, and translations receive an *equivalent* value in the *tipus* ‘type’ attribute.

Example 1.1. XML of the entry ‘administrador -a de xarxes’

```
<fitxa num="32">
  [...]
  <denominacio llengua="ca" tipus="principal" jerarquia="terme
    pral." categoria="n m, f">administrador -a de xarxes</
    denominacio>
  <denominacio llengua="ca" tipus="remissio" jerarquia="terme pral
    ." categoria="n m, f">gestor -a de xarxes</denominacio>
  [...]
  <denominacio llengua="en" tipus="equivalent" jerarquia="terme
    pral." categoria="">network administrator</denominacio>
  [...]
</fitxa>
```

The main goal of this work is to link both terminologies by merging the entries shared across them, thus enriching the information provided in one terminology with that of the other through the use of the semantics defined in the *lemon-ontolex* model.

4.2 URI naming strategy

The Terminoteca RDF combines two different URI naming strategies. On the one hand, the strategy adopted in the conversion of Terminesp to RDF, and, on the other, the one followed in the transformation of TERMCAT. Both terminologies differ in the way their original data were identified and they were converted independently of one another, which led to stick to similar but not entirely equal strategies. Terminesp source data included for each entry a numeric identifier which served as a unique concept identifier in the terminology and was propagated throughout the chains LexicalEntry – Lexical Sense – skos:Concept,

and `LexicalEntry` – Form. In fact, `Termine`sp URIs were constructed with identifier preservation in mind²⁰, e.g.: `lexiconEN/36995en-sense`. `TERMCAT`, in contrast, lacks these identifiers but provides an entry number unique in each of its domain lexica. The strategy in `TERMCAT` goes back to the one proposed in the conversion of the `Apertium Bilingual Dictionaries` to `RDF` (Gracia et al., 2016), where lexical entries, forms, and senses include in their URI the written representation of the words they describe along with their part of speech and language code (e.g.: `lexiconEN/accelerator-n-en`). In order to allow for their merge, the URIs of `Termine`sp lexical entries were changed to their counterpart in the second strategy, thus favouring reusability (i.e., there is no need to know the identifier of the lexical entry to reuse its URI but only its canonical written representation and its part of speech). The URIs of `Termine`sp senses, however, remain the same (i.e., preserve the identifiers).

4.3 Modelling

For the modelling of `Termine`sp and `TERMCAT` we build upon existing work on converting `Termine`sp to `RDF` (Bosque-Gil et al., 2015). Following that approach, each term is regarded as an `ontolex:LexicalEntry` and its meaning is represented with a `skos:Concept`. The relation between the lexical entry and the meaning is reified in the element `ontolex:LexicalSense`. Translations, terminological relations between a term and its scientific denomination and other possible semantic relations such as synonymy, the latter being present only in `TERMCAT` and not in `Termine`sp, are encoded at this level via `vartrans:Translation`, `vartrans:TerminologicalRelation` and `vartrans:SenseRelation` respectively.

`Termine`sp was first modelled on the basis of the concepts a term denotes. Polysemic terms were divided into as many lexical entries as senses a word had. This was captured by the URI naming strategy first mentioned above, which relied on concept identifiers (eg. `:lexiconES/63841es`). As an example, the word *red* ‘network’ in Spanish occurred three times as a lexical entry in the previous linked data version of `Termine`sp (Bosque-Gil et al., 2015), one for each concept it denotes: 38756, 54593, 63841. By transforming the URIs of the lexical entries according to the naming strategy devised for `TERMCAT`, these entries are then merged into one single lexical entry of the form `:lexiconES/red-n-es`, even though their senses still keep their identifiers as part of the URI and are mapped to different concepts.

In `TERMCAT` we were dealing with different dictionaries and not with a single terminology file as in `Termine`sp. The information regarding each term in a domain lexicon is considered to pertain to a specific sense of that entry in that domain. This results in a configuration similar to that of `Termine`sp: a single entry with a URI formed by the written representation, the part of speech and

²⁰ As recommended at <https://joinup.ec.europa.eu/community/semic/document/10-rules-persistent-uris>

the language code merges the different entries extracted from all domain lexica in which the term occurs as record.

Both the part of speech, which in Terminesp was added on top of the source data but was already provided in TERMCAT XML files, as well as the gender, are encoded by using LexInfo (Cimiano et al., 2011) properties (`lexinfo:partOfSpeech` and `lexinfo:gender`, respectively). TERMCAT data also includes abbreviations, initialisms, dismissed terms, and synonyms in addition to translations from Catalan to other languages. All translations, be they among full forms, initialisms or other abbreviated forms, were modelled with the `vartrans` module. The module represents translations as relations between lexical senses that are reified in the element `vartrans:Translation`.

As an example, the term *Electronic Programming Guide* has two different forms, one being the full form and the other one the initialism *EPG*. Thus, its lexical entry is linked to its full form by an `ontolex:canonicalForm` property and by its initialism via `ontolex:otherForm`. The initialism form is itself the lexical form of an independent lexical entry *EPG*, which has as term type `lexinfo:initialism` and which is linked to the lexical entry *Electronic Programming Guide* through `lexinfo:initialismFor`. This entry has a lexical sense which is mapped to the same `skos:Concept` as the full form of the term. Other abbreviated forms that are not initialisms are considered `lexinfo:AbbreviatedForm(s)`.

Synonyms have been modelled with the `vartrans` module and they constitute their own lexical entries in the lexicon. For any given entry that includes a synonym, the lexical sense of that synonym is defined as `vartrans:target` of a `vartrans:SenseRelation` with the sense of the first entry as `vartrans:source`. Both senses point to the same `skos:Concept`, which is domain-dependent. The fact that the relation is of synonymy is stated with the `vartrans:category` property pointing to the LexInfo property `lexinfo:synonym`.

4.4 Generation and Linking

The tool OpenRefine²¹ with its extension for linked data²² was used to generate the RDF and to link both terminologies. The linking is established at the level of lexical entries by merging Terminesp and TERMCAT terms through the use of the following pattern as URI naming strategy: `<written representation of term> - <part of speech> - <language code>`. The two images below illustrate the configurations of Terminesp and TERMCAT lexica before converting them to RDF (Figure 2) and at the current state (Figure 3).

Table 1 provides an overview of the number of terms that occur separately in Terminesp and TERMCAT Spanish, English and French lexica and the number of entries merged in this last version of the Terminoteca.

By adopting the URI naming strategy introduced above, Terminesp entries that were previously divided by concept are now merged into a single

²¹ <http://openrefine.org/>

²² The extension is hosted in GitHub:<https://github.com/sparkica/LODRefine>

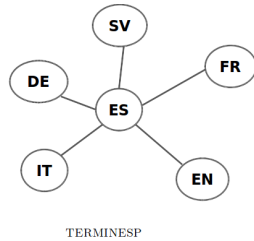


Fig. 2. Terminesp and TERMCAT before merging

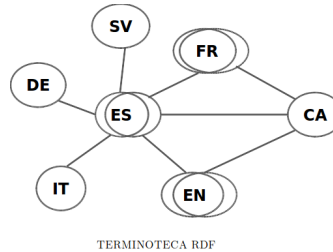
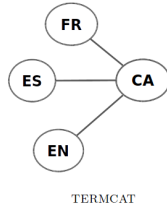


Fig. 3. Terminoteca RDF

	LexiconES	LexiconEN	LexiconFR
Exclusive from Terminesp	26941	14211	13970
Exclusive from TERMCAT	5583	6006	1028
Present in both (merged)	1087	878	297
Total	33611	21095	15295

Table 1. Distribution of entries in the Terminoteca

one if they share the written representation and part-of-speech. For instance, the entry `:lexiconES/ensayo-n-es` ‘test’ is now linked to eleven lexical senses (e.g. `:lexiconES/39026es-sense`, `:lexiconES/ensayo-n-es-IndustriaElectronica-sense`, etc.), ten of them extracted from Terminesp and with different definitions, and one of them coming from TERMCAT Electronics domain lexicon. Some of the senses are linked to the lexical senses of entries in other languages through a translation relation: `lexiconCA/assaig-n-ca` (from TERMCAT), and `:lexiconFR/essai-n-fr` (from Terminesp).

Likewise, one of the lexical senses of the entry `:lexiconES/cámara+de+televisión-n-es`²³ receives a synonym from TERMCAT (`:lexiconES/telecámara-n-es`), as well as a translation to Catalan `:lexiconCA/càmera+de+televisió-n-ca`. The English translation `:lexiconEN/television+camera-n-en` comes from both Terminesp and TERMCAT, and the links to the French `:lexiconFR/caméra+de+télévision-n-fr` and German entries `:lexiconDE/Fernsehkamera-n-de` are provided by Terminesp.

5 Conclusion and Future Work

We have presented the first steps towards the creation of Terminoteca RDF, a repository of linked terminologies that cover official languages in Spain and which in addition provide translations to other European languages. By relying on linked data technologies, terms formerly described in an isolated fashion, belonging even to the same terminology, are now reusable, linked to one another,

²³ We are not using IRIs, so the actual URI reads: `:lexiconES/c/C3%A1mara+de+televisi%C3%B3n-n-es`. Accents are included here for the sake of readability.

and enriched with the information provided by complementary terminological resources. The integration with other terminologies in official languages in Spain (Galician, Basque), as well as the linking to other available external resources such esDBpedia, Apertium dictionaries or IATE RDF are planned as continuation steps.

6 Acknowledgements

This work is supported by the Spanish Ministry of Economy and Competitiveness through the project 4V (TIN2013-46238-C4-2-R), the Excellence Network ReTeLe (TIN2015-68955-REDT), the Juan de la Cierva program and by the Spanish Ministry of Education, Culture and Sports through the Formación del Profesorado Universitario (FPU) program. We would also like to thank AENOR and AETER for providing us with the source data to initiate this effort.

References

- Bizer, C., T. Heath, and T. Berners-Lee (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)* 5(3), 1–22.
- Bosque-Gil, J., J. Gracia, G. Aguado-de Cea, and E. Montiel-Ponsoda (2015). Applying the OntoLex Model to a Multilingual Terminological Resource. In *The Semantic Web: ESWC 2015 Satellite Events*, pp. 283–294. Springer.
- Cimiano, P., P. Buitelaar, J. McCrae, and M. Sintek (2011). Lexinfo: A declarative model for the lexicon-ontology interface. *Web Semantics: Science, Services and Agents on the World Wide Web* 9(1), 29–51.
- Cimiano, P., J. McCrae, V. Rodríguez-Doncel, T. Gornostaya, A. Gómez-Pérez, B. Siemoneit, and A. Lagzdins (2015). Linked Terminology: Applying Linked Data Principles to Terminological Resources. In *Proceedings of eLex 2015*, pp. 504–517.
- Gracia, J., E. Montiel-Ponsoda, D. Vila-Suero, and G. Aguado-de Cea (2014). Enabling Language Resources to Expose Translations as Linked Data on the Web. In N. Calzolari, K. Choukri, T. Declerck, T. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *LREC 2014, Ninth International Conference on Language Resources and Evaluation*. ELRA European Language Resources Association.
- Gracia, J., M. Villegas, A. Gómez-Pérez, and N. Bel (2016). The Apertium Bilingual Dictionaries on the Web of Data. *Semantic Web Journal [submitted for peer review]*.
- Manola, F. and E. Miller (2004, February). RDF primer. Technical report, W3C Recommendation.
- Maslias, R. (2014). Combine EU Terminology with Communication and Ontology Research. In *Terminology and Knowledge Engineering 2014*, Softconf.org, pp. 49–57.

- Miles, A. and S. Bechhofer (2009). SKOS-Simple Knowledge Organization System Reference. Retrieved April 11, 2011.
- Navigli, R. and S. P. Ponzetto (2010). BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 216–225. Association for Computational Linguistics.
- Vila-Suero, D., A. Gómez-Pérez, E. Montiel-Ponsoda, J. Gracia, and G. Aguado-de Cea (2014, August). *Publishing Linked Data: the multilingual dimension*, pp. 101–118. Springer Berlin Heidelberg.

Terminology and ontology development in the domain of Islamic archaeology^{*}

Bruno Almeida,¹ Christophe Roche² and Rute Costa³

¹ NOVA CLUNL, Faculdade de Ciências Sociais e Humanas, Universidade NOVA de Lisboa,
Avenida de Berna, 26-C 1069-061 Lisboa, Portugal
`bruno.r.almeida@gmail.com`

² Condillac Research Group, LISTIC, Université de Savoie Mont Blanc, Campus Scientifique,
73376 Le Bourget du Lac, France
`christophe.roche@univ-savoie.fr`

³ NOVA CLUNL, Faculdade de Ciências Sociais e Humanas, Universidade NOVA de Lisboa,
Avenida de Berna, 26-C 1069-061 Lisboa, Portugal
`rute.costa@fcsb.unl.pt`

Abstract. This paper describes an example regarding the terminology of Islamic pottery artefacts in Portuguese and Spanish in the context of an ongoing Ph D project. The approach followed in this paper places knowledge representation at the core of terminology work. More specifically, the development of an ontology, i.e. a formal and computational conceptualisation, enables the integration of a multilingual termbase in the semantic web as linked data, targeted at experts and students of archaeology. This approach allows for the preservation of linguistic diversity, as reflected by the different linguistic practices engaged by Portuguese and Spanish archaeologists in scholarly communication.

Keywords: terminology, knowledge representation, ontologies, multilingual termbases, semantic web, pottery artefacts, Islamic archaeology

1 Terminology and Islamic archaeology: the case of pottery artefacts

Islamic presence in the Iberian Peninsula covered a period of nearly eight centuries (from 711 to 1492 A.D.), and left behind a wide range of materials, such as pottery, architectural fragments, weaponry, jewellery and glassware. For many decades, archaeologists in Portugal and Spain have worked on the description, analysis and comparison of these objects, focusing on properties such as function, shape, materials, manufacturing and decorative techniques.

^{*} This research has been financed by Portuguese National Funding through the FCT – Fundação para a Ciência e a Tecnologia as part of the project Centro de Linguística da Universidade Nova de Lisboa - UID/LIN/03213/2013.

Pottery is considered to be one of the most important types of artefacts for archaeologists, not only because of its high durability but also due to its cultural significance. According to Kipfer, pottery is “often one of the clearest indicators of cultural differences, relations and developments” (Kipfer, 2000, p. 452). Since the date of manufacture usually can be determined, pottery sherds are also important in dating other finds (*ibid.*). In the last decades, the study of Islamic pottery in Portugal and Spain has furthered the understanding of the culture and society of the al-Andalus:¹ its eating habits, everyday life, trade relations, technical development and even its symbolism and ideology (Gómez Martínez, 2004).

A significant part of this knowledge is only made possible by the typological analysis of pottery artefacts, which enables the comparison and study of related finds. Within archaeology, ‘typology’ is defined as “the classification of objects, structures, or specimens by subdividing observed populations into a theoretical sequence or series of groups (types) and subgroups (subtypes) according to consideration of their qualitative, quantitative, morphological, formal, technological, and functional attributes.” (Darvill, 2009).

In Portugal, the lack of terminology harmonisation has been referred in the past as a hurdle in scholarly communication in the domain of Islamic archaeology (Torres, Gómez Martínez, & Ferreira, 2003). Furthermore, terminology work is seen as a means to acquire and organise expert knowledge in this domain (*ibid.*). In recent years, the need to revitalise the studies on Islamic pottery in Portugal has led to the creation of the CIGA research group (*Cerâmica Islâmica do Gharb al-Ândalus*), which presently consists of twelve archaeologists.² The focus of this group was the creation of a shared database describing the most representative instances of Islamic pottery in the Gharb al-Andalus³ (Bugalhão et al., 2010). Underlying the creation of this database is a common typology and terminology of artefacts, shapes, and manufacturing and decorative techniques. CIGA’s typology of artefacts is based on eight classes, according to the theoretical purpose of the objects, namely: (i) storage and transportation, (ii) kitchenware, (iii) tableware, (iv) lighting objects, (v) household objects, (vi) agricultural and handicraft objects, (vii) recreational and ritual objects and (viii) construction materials. Each class is further divided into subclasses according to the formal attributes of the objects. Furthermore, definitions or descriptions in natural language are provided for each subclass, as well as graphical representations in the form of archaeological illustrations (Bugalhão et al., 2010).

The importance of terminology in archaeology, as evidenced by the CIGA group, raises several questions of import to our project, which is centred on the creation of a multilingual termbase in the domain and its integration in the semantic web. In this paper we will focus on the formal and conceptual analysis of Islamic pottery artefacts, following an interdisciplinary approach to terminology. This approach places knowledge representation at the core of terminology work, following previous work

¹ ‘Al-Andalus’ refers to the territory of the Iberian Peninsula and Septimania under Islamic occupation.

² More information available at <http://www.camertola.pt/info/ciga>.

³ Western region of the Iberian Peninsula under Islamic rule, which roughly corresponds to the continental territory of present day Portugal.

in the framework of ontoterminology (Roche, 2007). More specifically, we will show how an ontology may represent a language independent conceptualisation,⁴ allowing for the operationalisation of a multilingual termbase meant for experts and students of archaeology.

The example presented in this paper was drawn from the analysis of several texts (quoted below) written by Portuguese and Spanish archaeologists, including relevant graphical information. It should be noted that our conclusions may change as new data is gathered. Translations and equivalent designations in English are provided in order to facilitate communication in this paper.

2 Modelling artefact types: the case of lighting objects in pottery

In the typology of the CIGA group, the class of ‘lighting objects’ is divided into the subclasses referred to by the Portuguese terms *candil*, *candeia*, *candeia de pé* and *lanterna*. *Candil* is defined as a “lighting object with closed chamber”, while *candeia* is defined as a “lighting object with an open chamber”.⁵ *Candeia de pé* is defined as a “lighting object with an open chamber supported by a high foot”. Finally, *lanterna* is described as a “closed form with a globular body and central orifice, used for lighting in open spaces”.⁶ Fig. 1 illustrates representative instances of the named subclasses of pottery lighting objects.

The available information leads us to infer that *candeia de pé* is actually a subclass of *candeia*, since *candeia de pé* is a lighting object with an open chamber, with the delimiting characteristic of ‘being supported by a high foot’.⁷ We also infer that the type of object depicted in Fig. 1-II differs from *candeia de pé* by having a flat base instead of a high foot. While *candeia* and *candil* are clearly defined, being distinguished by the configuration of the chamber (open or closed), *lanterna* is described by typical characteristics (i.e. ‘globular body with a central orifice’) and the more specific purpose of lighting in open spaces. We propose that *candeia* and *candil* should belong to a subclass of lighting objects devised for lighting in closed spaces

⁴ By ‘language independent conceptualisation’ we mean a concept system that is not bound by any particular natural language.

⁵ Following Rice, an open vessel is generated by an unrestricted orifice, whose “diameter is equal to or greater than the maximum diameter of the body” (Rice, 2015, ch. 13.4.3.1). On the other hand, a closed vessel is generated by a restricted orifice.

⁶ According to this information, *candil* can be referred to in English as ‘closed lamp’, *candeia* as ‘open lamp’, *candeia de pé* as ‘foot lamp’ and, finally, *lanterna* as ‘lantern’. We should note that these terms may also refer to Islamic artefacts made of other materials besides pottery, which will not be covered in this paper.

⁷ According to the ISO terminology standards, a delimiting characteristic is an essential characteristic used for distinguishing between related concepts. By ‘essential characteristic’ we mean a characteristic that is essential in understanding a concept, which highlights its cognitive nature (ISO 1087-1:2000).

(which we can refer to as ‘lamp’ in English), since only *lanterna* has the purpose of providing a light source in open spaces.



Fig. 1. Archaeological illustration of the class of ‘lighting objects’ according to the CIGA group. From left to right: I. *candil*, II. *candeia*, III. *candeia de pé*, IV. *lanterna* (Source: Bugalhão *et al.*, 2010, p. 471).

Regarding the Spanish sources, Rosselló-Bordoy defines *candil* as a “portable or fixed element for domestic lighting” (Rosselló-Bordoy, 1991, p. 174), which corroborates our analysis that the lamp is an object meant for closed spaces. In his earlier work, Rosselló-Bordoy distinguished between several formal variants of *candil*, consisting essentially on the types of artefacts depicted in Fig. 1, I-III (Rosselló-Bordoy, 1978, pp. 48-55). These variants include a subclass referred to in Spanish as *candil de pie alto* (which has an open vessel, similar to Fig. 1, III), four closed variants (depending on the geometrical shape of the chamber), and an open variant without a foot (similar to Fig. 1, II). Therefore, the Spanish term *candil* denotes any type of lamp (open or closed). Rosselló-Bordoy also lists *fanal* or *linterna* within the class of lighting objects, corresponding to the same type of artefact depicted in Fig. 1, IV.⁸ Gómez Martínez provides a definition of *fanal* in line with the CIGA group: “*fanal* or *linterna* is defined as a closed form inside which fire is contained for the purpose of lighting in open spaces” (Gómez Martínez, 2004, p. 278). In the case of lamps, Spanish archaeologists also use the terms *candil de piquera* (‘nozzled lamp’) and *candil de pellizco* (‘pinched lamp’). These terms refer to the shape of the beak of these objects, which either have a nozzle or a pinched beak meant for holding a wick. However, ‘nozzled lamp’ and ‘pinched lamp’ refer to the same objects as ‘closed lamp’ and ‘open lamp’, respectively, as can be observed in the examples represented in Fig. 1. This is evidenced by Navarro Palazón and Jiménez Castillo (2007), who use the terms *candil de piquera* and *candil de pellizco* as synonyms of *candil de cazoleta cerrada* and *candil de cazoleta abierta*, respectively.⁹

⁸ Rosselló-Bordoy also includes *almenara* in the class of lighting objects, which is defined as “a sort of multiple *candil* or support for holding several *candiles*” (Rosselló-Bordoy, 1991, p. 174). However, this object seems to be ill-defined, as its existence is only documented in metal and not in pottery (Gómez Martínez 2004, p. 277).

⁹ These authors also corroborate our analysis that ‘foot lamp’ is a type of ‘open lamp’: “Durante los primeros siglos en al-Andalus se empleó un *candil*, denominado genéricamente *de piquera* o *de cazoleta cerrada*, derivado de las lucernas clásicas. Hacia la segunda mitad del siglo XII llegan a la Península Ibérica, desde el Mediterráneo oriental, dos nuevos tipos, llamados *de cazoleta abierta* o *pellizco* [...] y *de pie alto* [...], este último es básicamente un *candil de pellizco* dotado de una *peana*.” [our emphasis] (Navarro Palazón & Jiménez Castillo, 2007, p. 312).

Our analysis leads to the concept system represented in the UML class diagram shown in Fig. 2, based on the principle of genus and specific difference.¹⁰ Concepts are labelled with identifiers in English.¹¹

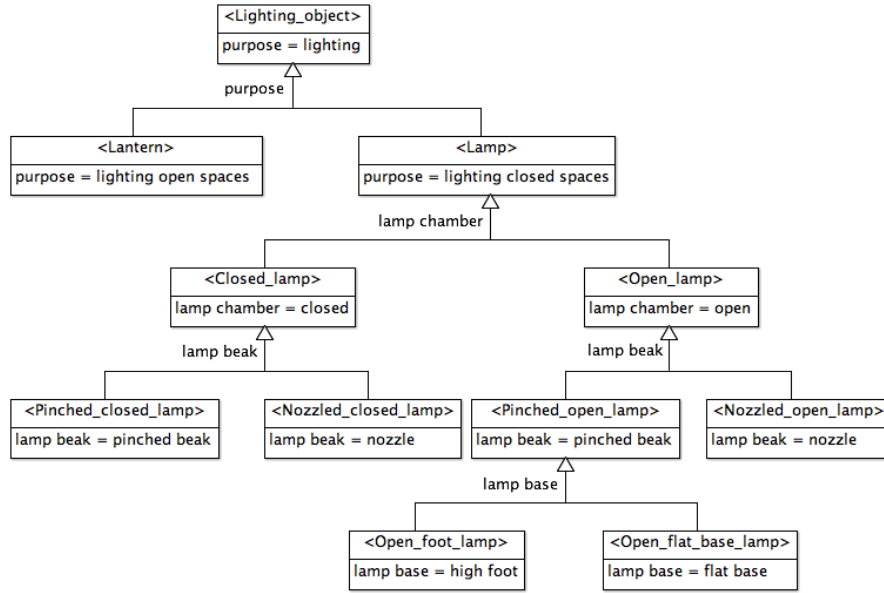


Fig. 2. Concept system of the class of 'lighting objects'.

This concept system can be used as a basis for an ontology of lighting objects. The following axioms provide an ontological definition – i.e. a formal, constructive definition (Roche, 2015) – of the relevant concepts in our ontology:¹²

$$\text{Lighting} \equiv \{\text{lighting_open_spaces}\} \cup \{\text{lighting_closed_spaces}\} \quad (1)$$

$$\text{Lighting_object} \equiv \text{Islamic_pottery_artefact} \cap \exists \text{hasPurpose.Lighting} \quad (2)$$

$$\text{Lantern} \equiv \text{Lighting_object} \cap \exists \text{hasPurpose.}\{\text{lighting_open_spaces}\} \quad (3)$$

$$\text{Lamp} \equiv \text{Lighting_object} \cap \exists \text{hasPurpose.}\{\text{lighting_closed_spaces}\} \quad (4)$$

$$\text{Lighting_object} \sqsubseteq \text{Lantern} \cup \text{Lamp} \quad (5)$$

¹⁰ This principle was followed because not only is it consistent with the available data, but also due to its usefulness in producing a conceptualisation in line with the ISO standards on terminology.

¹¹ Concept identifiers, which are only relevant to identify units of knowledge in a conceptualisation, are represented between angle brackets to further distinguish them from terms (Roche, 2012).

¹² <Pinched_closed_lamp> and <Nozzled_open_lamp> are not defined because they do not have any instances in Islamic pottery. Therefore, in this domain, a <Closed_lamp> is always a <Nozzled_closed_lamp> and an <Open_lamp> is always an <Pinched_open_lamp>.

$$\begin{aligned}
& \text{Lantern} \sqcap \text{Lamp} \sqsubseteq \perp & (6) \\
& \text{Lamp_chamber} \equiv \{\text{open}\} \sqcup \{\text{closed}\} & (7) \\
& \text{Open_lamp} \equiv \text{Lamp} \sqcap \exists \text{hasLampChamber}.\{\text{open}\} \sqcap \forall \text{hasLampChamber}.\{\text{open}\} & (8) \\
& \text{Closed_lamp} \equiv \text{Lamp} \sqcap \exists \text{hasLampChamber}.\{\text{closed}\} \\
& \quad \sqcap \forall \text{hasLampChamber}.\{\text{closed}\} & (9) \\
& \text{Lamp} \sqsubseteq \text{Open_lamp} \sqcup \text{Closed_lamp} & (10) \\
& \text{Open_lamp} \sqcap \text{Closed_lamp} \sqsubseteq \perp & (11) \\
& \text{Lamp_beak} \equiv \{\text{pinched}\} \sqcup \{\text{nozzle}\} & (12) \\
& \text{Pinched_open_lamp} \equiv \text{Open_lamp} \sqcap \exists \text{hasLampBeak}.\{\text{pinched}\} \\
& \quad \sqcap \forall \text{hasLampBeak}.\{\text{pinched}\} & (13) \\
& \text{Nozzled_closed_lamp} \equiv \text{Closed_lamp} \sqcap \exists \text{hasLampBeak}.\{\text{nozzle}\} \\
& \quad \sqcap \forall \text{hasLampBeak}.\{\text{nozzle}\} & (14) \\
& \text{Lamp_base} \equiv \{\text{high_foot}\} \sqcup \{\text{flat_base}\} & (15) \\
& \text{Open_foot_lamp} \equiv \text{Pinched_open_lamp} \sqcap \exists \text{hasLampBase}.\{\text{high_foot}\} \\
& \quad \sqcap \forall \text{hasLampBase}.\{\text{high_foot}\} & (16) \\
& \text{Open_flat_base_lamp} \equiv \text{Pinched_open_lamp} \sqcap \exists \text{hasLampBase}.\{\text{flat_base}\} \\
& \quad \sqcap \forall \text{hasLampBase}.\{\text{flat_base}\} & (17) \\
& \text{Pinched_open_lamp} \sqsubseteq \text{Open_foot_lamp} \sqcup \text{Open_flat_base_lamp} & (18) \\
& \text{Open_foot_lamp} \sqcap \text{Open_flat_base_lamp} \sqsubseteq \perp & (19)
\end{aligned}$$

Delimiting characteristics are represented by roles whose range is specified by individual values. These values belong to the concepts defined in axioms (1), (7), (12) and (15). The covering and disjointness axioms required for this conceptualisation are defined in (5), (6), (10), (11) and (18), (19).

Ontologies allow for the integration of multilingual resources in the semantic web, functioning as their conceptual and computational underpinning. The question now arises regarding the specificity of each language. This will be addressed in the following chapter.

3 The terminology of lighting objects in Portuguese and Spanish

Although we assume that concepts are extra-linguistic constructs, it does not entail that terminology is independent from the linguistic practices engaged by domain experts in scholarly communication. Terms are determined by cultural and linguistic factors (Lerat, 1995), which makes them more than mere labels for concepts: they are,

in fact, lexical items in their own right, acquiring their status by virtue of their usage and recognition within a specialised community of practice.

Turning our attention to the example at hand, it is clear that there is some difference between both languages. In Portuguese, there does not seem to be a suitable term for <Lamp>, as the archaeologists use the more specific terms *candeia* and *candil*. There is, however, evidence in Portuguese texts that the concepts denoted by these terms are closely related.¹³ Nevertheless, <Lamp> should remain an unnamed concept in this language in our termbase in order to reflect the specificity of the linguistic practices of Portuguese archaeologists. Fig. 3 represents the Portuguese terminology of lighting objects according to the data available at this time.¹⁴

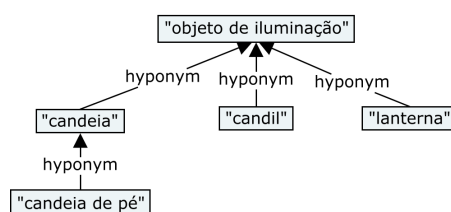


Fig. 3. The terminology of lighting objects in Portuguese.

In the case of Spanish, every concept is denoted by at least one term in scholarly communication, including three notable cases of synonymy. The information regarding the Spanish terminology is represented in Fig. 4.

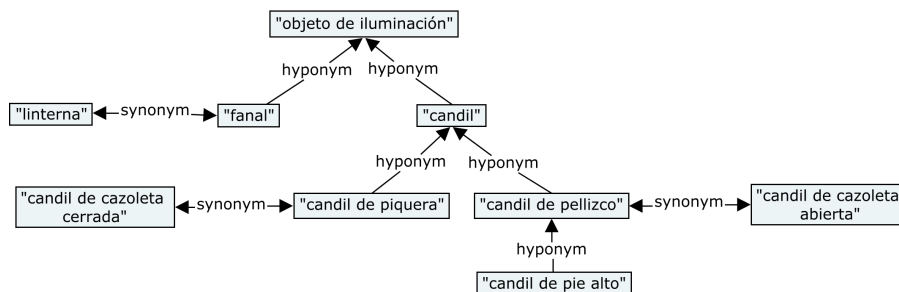


Fig. 4. The terminology of lighting objects in Spanish.

As we can see, neither of these lexical networks, which represent language specific information, is isomorphic to the concept system outlined in Fig. 2, which represents knowledge shared within a community of practice. From an onomasiological point of

¹³ For instance: “[...] distinguindo deste modo o CANDIL, de forma fechada, da CANDEIA que seria a forma aberta que se manteve praticamente até aos nossos dias” [emphasis in the original] (Torres, Gómez Martínez & Ferreira, 2003, p. 129).

¹⁴ Terms are represented between double quotation marks (Roche 2012). The lexical networks in this section are based on the relations of hyponymy (generic-specific relation between term meanings) and synonymy (relation of equivalence between term meanings).

view, the concepts in our ontology are denoted by the following terms in each language:

- <Lighting_object> *isDenotedBy* “objeto de iluminação” (pt), “objeto de iluminación” (es);
- <Lantern> *isDenotedBy* “lanterna” (pt), “fanal” (es), “linterna” (es);
- <Lamp> *isDenotedBy* “candil” (es);
- <Closed_lamp> (and <Nozzled_closed_lamp>) *isDenotedBy* “candil” (pt), “candil de piquera” (es), “candil de cazoleta cerrada” (es);
- <Open_lamp> (and <Pinched_open_lamp>, <Open_flat_base_lamp>) *isDenotedBy* “candeia” (pt), “candil de pellizco” (es), “candil de cazoleta abierta” (es);
- <Open_foot_lamp> *isDenotedBy* “candeia de pé” (pt), “candil de pie alto” (es).

The interface between our termbase and the ontology described in the last section can be achieved by adapting a model such as Lemon (Lexicon Model for Ontologies), which is under development by the W3C Ontology-Lexica Community Group.¹⁵ This also facilitates the access to the termbase as linked data in RDF (Resource Description Framework). To give an example, the following RDF code in Turtle syntax represents a terminological entry for “candil” in Portuguese, referring to the concept <Nozzled_closed_lamp>.¹⁶

```
:candil-pt a ontolex:LexicalEntry, ontolex:Word ;
  ontolex:canonicalForm :candil-pt#CanonicalForm ;
  rdfs:label "candil"@pt ;
  ontolex:language "pt" ;
  ontolex:sense :candil-pt#Sense .

:candil-pt#CanonicalForm a ontolex:Form ;
  ontolex:writtenRep "candil"@pt .

:candil-pt#Sense a ontolex:LexicalSense ;
  ontolex:reference <http://.../Nozzled_closed_lamp> ;
  skos:definition "Objeto cerâmico de origem islâmica para iluminação doméstica com depósito fechado e bico de canal."@pt .

:senseRelation a vartrans:SenseRelation ;
  vartrans:source :objeto_de_iluminacao-pt#Sense ;
  vartrans:target :candil-pt#Sense ;
  vartrans:category :hyponym .
```

This approach enables the full integration of an archaeology termbase in the semantic web. This facilitates the access to, and manipulation of, terminological and

¹⁵ More information available at <https://www.w3.org/community/ontolex/>.

¹⁶ We should note that in this example “candil” refers only to a subclass of pottery artefacts. However, further senses of the term can be defined in order to provide a more complete account of its meaning within Islamic archaeology.

conceptual data by both human and machine agents, which is paramount in the context of information society, allowing for a more efficient construal of knowledge.

4 Concluding remarks

The analysis outlined in this paper is only possible by following an interdisciplinary approach to terminology, looking beyond linguistics and specialised lexicography. This has an important precedent in the work of Wüster (1979), for whom terminology theory overlaps with logic, ontology and information science.

Terminology as a domain emerges from the interaction between disciplines centred on the study of language and knowledge, from which it derives its principles and methods as a discipline. Presently, the object of study of terminology is recognised as being multidimensional and, therefore, irreducible to any particular discipline (Cabr , 2000). We assume that terminology has fundamentally a double dimension: linguistic and conceptual (Costa, 2013; Roche 2015; Santos & Costa, 2015). While the linguistic dimension pertains to terms, their behaviour in discourse and their role within specialised communities of practice, the conceptual dimension consists on the knowledge shared within these communities and how it can be represented for multiple applications (computational or otherwise). Indeed, the core elements of terminology remain the concept (unit of knowledge), the term (specialised lexical item), and the relationship between these elements, in which lies the specificity of terminology as a domain at the crossroads between language and knowledge (Costa, 2013).

In the past decades, terminology has been characterised by a “plurality of theoretical approaches” (Costa, 2006) in which linguistics plays an increasingly dominant role and thereby relegating terminology to a sort of specialised lexicography. However, the need for the operationalisation of multilingual terminology resources, i.e. their computational representation, requires an approach in line with knowledge representation, a field of artificial intelligence, which once again brings into question the need to widen the scope of terminology as an interdisciplinary domain. This opens up important applications for the discipline in the context of information society, from computer assisted translation to SEO, semantic search engines and interactive navigation tools in data repositories (Roche, 2015).

This paper focused on the conceptual dimension of terminology. We saw how lexical networks, which represent language specific information, are not isomorphic to a concept system, which represents shared knowledge in the domain. Placing ontology development at the core of terminology work enables the operationalisation of multilingual terminologies in the semantic web, allowing for the description of the linguistic diversity manifested in scholarly communication.

References

- Bugalh o, J. *et al.* (2010). CIGA: projecto de sistematiza  o para a cer mica isl mica do Gharb al- ndalus. *Xelb*, 10, 455-476.

- Cabré, M. T. (2000). Terminologie et linguistique : la théorie des portes. *Terminologies nouvelles*, 21, 10-15.
- Costa, R. (2006). Plurality of theoretical approaches to terminology. In: H. Picht (Ed.), *Modern approaches to terminological theories and applications* (pp. 79-89). Bern: Peter Lang.
- Costa, R. (2013). Terminology and specialised lexicography: two complementary domains. *Lexicographica*, 29, 29-42.
- Darvill, T. (2009). *The concise Oxford dictionary of archaeology* (Online ed.). Oxford: Oxford University Press.
- Gómez Martínez, S. (2004). *La cerámica islámica de Mértola: producción y comercio*. Madrid: Universidad Complutense.
- ISO 1087-1 (2000). Terminology work – Vocabulary – Part 1: Theory and application. Geneva: ISO.
- Kipfer, B. A. (2000). *Encyclopedic dictionary of archaeology*. New York: Springer Science+Business Media.
- Lerat, P. (1995). *Les langues spécialisées*. Paris: Presses universitaires de France.
- Navarro Palazón, J. & Jiménez Castillo, P. (2007). *Siyasa: estudio arqueológico del despoblado andalusí (ss. XI-XIII)*. Granada: Escuela de Estudios Árabes de Granada.
- Rice, P. M. (2015). *Pottery analysis: a sourcebook*. (Second ed.). (Kindle ed.). Chicago: The University of Chicago Press.
- Roche, C. (2012). Ontoterminology: how to unify terminology and ontology into a single paradigm. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)* (pp. 2626–2630). Paris: ELRA.
- Roche, C. (2007). Terme et concept : fondements pour une ontoterminologie. In: TOTh 2007 (pp. 1-22). Annecy: Institut Porphyre.
- Roche, C. (2015). Ontological definition. In: H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology: vol. 1* (pp. 128–152). Amsterdam: John Benjamins Publishing Company.
- Santos, C. & Costa, R. (2015). Domain specificity: semasiological and onomasiological knowledge representation. In: H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology: vol. 1* (pp. 153–179). Amsterdam: John Benjamins Publishing Company.
- Rosselló-Bordoy, G. (1978). *Ensayo de sistematización de la cerámica árabe en Mallorca*. Palma de Mallorca: Institut d'Estudis Baleàrics.
- Rosselló-Bordoy, G. (1991). *El nombre de las cosas en al-Ándalus: una propuesta de terminología cerámica*. Palma de Mallorca: Museo de Mallorca.
- Torres, C., Gómez Martínez, S., & Ferreira, M. B. (2003). Os nomes da cerâmica medieval: inventário de termos. In: *Actas das 3as Jornadas de Cerâmica Medieval e Pós-Medieval* (pp. 125-134). Tondela: Câmara Municipal.
- Wüster, E. (1979). *Introduction to the General Theory of Terminology and Terminological Lexicography*. Vienna: Springer.

LESS Can Indeed Be More:

Linguistic and Conceptual Challenges in the Age of Interoperability

Sara Carvalho¹²³, Rute Costa²³, Christophe Roche³²

¹ School of Technology and Management – University of Aveiro
R. Comandante Pinho e Freitas, 28 3750-127 Águeda – Portugal
`sara.carvalho@ua.pt`

² NOVA CLUNL – Faculty of Social Sciences and Humanities – Univ. NOVA de Lisboa
Av. de Berna, 26-C 1069-061 Lisboa – Portugal
`rute.costa@fcsh.unl.pt`

³ Condillac Research Group – LISTIC – Université de Savoie Mont Blanc
Campus Scientifique 73376 Le Bourget du Lac – France
`christophe.roche@univ-savoie.fr`

Abstract. The advent of the Semantic Web and, more recently, of the Linked Data initiative, has paved the way for new perspectives and opportunities in Terminology, namely regarding the operationalization of terminological products. Within the biomedical domain, changes have been substantial in the past decades and at their heart stand the current challenges regarding the production, use, storage and dissemination of medical data, information, and knowledge. In a context where biomedical terminological resources are becoming increasingly concept-oriented, terminology work should reflect a double dimension (both linguistic and conceptual) that may, in turn, support the aspired operationalization and interoperability in this field. Therefore, the purpose of this paper is to present a case study, based around the concept of <Laparoendoscopic single-site surgery>, in which a methodology anchored in Terminology’s double dimension aims to contribute to the enrichment of the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT).

Keywords. terminology’s double dimension; interoperability; laparoendoscopic single-site surgery; SNOMED CT.

1 Introduction

The advent of the Semantic Web and, more recently, of the Linked Data initiative, has paved the way for new perspectives and opportunities in Terminology, namely in what concerns the operationalization of terminological products. The increasingly collaborative work involving Terminology and ontologies – in the sense of Knowledge Engineering (KE) – has led to the development of numerous resources in several areas of knowledge, one of them being Medicine.

Within the biomedical domain, changes have been substantial in the past decades: on the one hand, health care provision has become more technology-based, with computerized examinations, procedures, prescriptions and health records. Furthermore, growing digital literacy has brought the patients into the driver's seat, where they have been playing a more active – and empowered – role. On the other hand, ageing population and the dramatic decline of the old-age support ratio have contributed to more pressure on public health expenditure, leading to the existing debate around the sustainability of social security systems and their role in health care. At the heart of all these issues stand the current challenges regarding the production, use, storage and dissemination of medical data, information, and knowledge. The ability to provide secure, reliable, efficient and cost-effective ways to process and exchange clinical information among all the stakeholders has emerged as the cornerstone of eHealth initiatives worldwide, with interoperability as one of the key elements¹.

In a context where biomedical terminological resources are becoming increasingly concept-oriented, it is of paramount importance for terminology work to reflect a double dimension (both linguistic and conceptual) that may support the aspired operationalization and interoperability in this field. It is believed that Terminology's input in the representation, organization, dissemination and, therefore, in the stabilization of specialized knowledge should be taken into account.

Hence, the purpose of this paper is to present a case study, based around the concept of <Laparoscopic single-site surgery>², in which a methodology anchored in Terminology's double dimension aims to contribute to the enrichment of a particular biomedical terminological resource: the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT). The concept under analysis integrates the wider scope of the EndoTerm project, presented in previous papers³. This article will thus be structured as follows: Section 2 will provide a brief overview of the theoretical background to this case study; Section 3 will be dedicated to SNOMED CT, particularly its logical and concept models; Section 4 will focus on the case study around the concept of <Laparoscopic single-site surgery>⁴, followed by some concluding remarks.

2 Two sides of the same coin: Terminology's double dimension

As mentioned above, the double dimension approach, which comprises both a linguistic and a conceptual dimension that are interrelated, has been described by Roche *et al.* (2009), Roche (2012, 2015), Costa (2013), and by Santos and Costa (2015). According

¹ Cf. the European Commission's eHealth Action Plan 2012-2020, available at <https://ec.europa.eu/digital-single-market/en/eu-policy-ehealth> (01.03.2016), or the World Health Organization's projects on eHealth (<http://www.who.int/ehealth/programmes>) (01.03.2016).

² A type of surgical procedure that is becoming more and more prevalent in several medical specialties. It is also known as LESS surgery.

³ Cf. Carvalho, Roche, and Costa (forthcoming); Carvalho, Roche, and Costa (2015).

⁴ Throughout this paper, concepts will be capitalized and written between single chevrons, whereas terms will be presented in lower case and between double quotation marks (cf. Roche 2015)

to Roche (2015: 136), Terminology is “both a science of objects and a science of terms”. For Costa (2013), it is precisely this double dimension, as well as the study of the relationship between one and the other, that grants Terminology its place as an autonomous scientific subject.

This double dimension perspective implies, therefore, that both the experts’ conceptualizations of a given subject field and the discourses produced by them must be taken into account in terminology work, thus leading to a complementarity of two fundamentally different dimensions. Consequently, both specialized texts and expert collaboration constitute invaluable resources in terminological work, provided that there is a supporting theoretical and methodological framework that allows the terminologist to maximize the potential within each dimension, and mostly of the synergies resulting from their interaction.

3 Current biomedical terminological resources and interoperability: the example of SNOMED CT

Medicine is currently undergoing significant challenges regarding the way clinical information and knowledge are produced, used, stored and shared. In recent years, many biomedical terminological resources have been designed or redesigned in order to incorporate ontology-based elements, thus evolving from “simple code-name-hierarchy arrangements, into rich, knowledge-based ontologies of medical concepts” (Cimino 2001). Concept-orientation has become one of the key principles of current biomedical resources and was, in fact, one of the twelve desiderata that, according to Cimino (1998), should support biomedical terminological systems in the 21st century.

One of these resources is SNOMED CT, currently owned and distributed by the International Health Terminology Standards Development Organization (IHTSDO). It is a comprehensive, multilingual health care terminology that due to its description-logic basis, supports the representation of clinical content in electronic health formats (namely Electronic Health Records – EHRs) in a consistent, reliable and computer-readable way⁵.

This resource has been built around three main components: the **concepts**, which represent clinical meanings, are organized into hierarchies, ranging from general to specific; the **descriptions**, which provide the human readable form of a concept, comprise the Fully Specified Name (FSN), representing “a unique, unambiguous description of a concept’s meaning”⁶, and the synonym (SYN). Each concept may have multiple synonyms, but only one is marked as “preferred” in a given language, whereas the remaining synonyms are marked as “acceptable”; finally, the **relationships**, which connect concepts to other related concepts, are used to logically define the meaning of a concept in a computer-processable way. There are two main types of relationships: subtype, or is_a relationships, which form the basis of SNOMED CT’s hierarchies, and attribute relationships, that associate the source concept (e.g. [abscess of heart]) with the value

⁵ For more information, see <http://www.ihtsdo.org/snomed-ct/what-is-snomed-ct>

⁶ Cf. IHTSDO (2014: 14–17).

of a defining characteristic. The characteristic (attribute) is specified by the relationship itself (e.g. |finding site|) and the value is provided by the destination concept (e.g. |heart structure|). Each one of these three components (concepts, descriptions and relationships) has its own unique numeric identifier.

SNOMED CT has been selected as the basis of our proposal for several reasons: **firstly**, the concept-based nature of this resource and the fact that it aims to integrate the linguistic and the conceptual, while preserving their fundamental differences, appear to be consistent with Terminology's double dimension and its core principles⁷; **secondly**, SNOMED CT's structure allows post-coordination, i.e. more complex concepts may be created from a set of more primitive components. In this type of system, also called compositional (cf. Coiera 2015; Duclos *et al.* 2014), there is no need to create all the elements in advance, but rather to ensure that all the basic building blocks exist. It is therefore possible to represent a given clinical content even when the precise concept is not present in SNOMED CT. This representation may occur via a standard compositional grammar that is both human-readable and computer-processable, thus enabling interoperability⁸. Moreover, this compositional approach to concept representation requires the definition of a set of logical rules (constraints)⁹ that will govern the way concepts and relationships can be combined, in order to prevent nonsense representations; **thirdly**, and unlike other resources of its kind, SNOMED CT is not limited to hierarchical concept relations; **finally**, the concept under analysis does not exist in this terminological system, so it is believed this proposal could contribute to enrich SNOMED CT's content.

4 LESS: a Brave New World for surgery?

As stated above, the development of the EndoTerm project, which was described in depth in Carvalho, Roche, and Costa (2015) and which aims at the creation of a multi-lingual terminological resource based around the concept of <Endometriosis>, led to the study of single port surgery, a relatively recent type of minimally invasive surgery. The further analysis of the concept pointed towards a lack of terminological consensus among the expert community, with a plethora of terms coined by individual groups and organizations. In fact, more than 20 have been identified in the literature¹⁰.

⁷ As well as with Ontoterminology (cf. Roche *et al.* (2009); Roche (2012); Roche (2015); Carvalho, Roche, and Costa (2015).

⁸ A concrete example of this compositional grammar will be presented in the next section.

⁹ Also known as the categorial structure (cf. ISO 17115: 2007). For example, `has_site` should occur only between concepts related to morphology and concepts referring to topography (e.g. `pyelonephritis is_a infection (morphology: -itis) which "has_site" kidney (topography: -nephro-)`).

¹⁰ Due to space constraints, it was not possible to include a table with all the collected designations in this paper (a total of 22). However, they can be found in Box *et al.* (2008); Gill *et al.* (2010); Autorino *et al.* (2011); Ramesh, Vidyashankar, and Dimri (2014); Georgiou *et al.* (2012); Springborg and Fader (2015); Escobar and Falcone (2014), just to name a few.

In order to solve this terminological dispersion, in 2008 a multidisciplinary medical consortium¹¹ decided to standardize the terminology in the field and proposed the term “laparoendoscopic single-site surgery” (also known as “LESS surgery”) as the one that most accurately depicted this surgical procedure. In addition, LESSCAR’s White Paper highlighted the characteristics that the concept should encompass: 1) a single entry port (or incision); 2) applicability to multiple locations (abdominal, thoracic or pelvic); 3) umbilical or extraumbilical access; 4) type of surgery (laparoscopic, endoscopic, or robotic); 5) type of surgical approach (percutaneous intraluminal and percutaneous transluminal). The group also required that all scientific publications on LESS surgery should include a “mandatory descriptive second line”, with details about the number and type of ports used, the type of laparoscope used, and the type of instruments used. (Gill *et al.* 2010).

Within the scope of the EndoTerm project, the gathered data concerning LESS surgery provided sufficient ‘food for thought’ in order to constitute a terminological case study. A number of questions interconnecting the linguistic and conceptual dimensions arose from the data analysis: i) are all the gathered terms actual synonyms from a terminological standpoint, i.e. representing the same concept and being interchangeable in all contexts (cf. ISO 1087-1, 2000)? ii) what usage has the expert community in the field of Gynecology¹² been making of these designations? iii) knowing that texts do not contain concepts themselves, but the linguistic usages of the terms that designate them, what type of information could be extracted from a set of natural language definitions? And in what way would that match a concept map validated by subject field experts? iv) given that the concept of <Laparoendoscopic single-site surgery> does not exist in SNOMED CT, what additional information would be necessary for its inclusion in this resource and how could it be represented in a way that enables interoperability?

In the literature, the designations used to refer to LESS surgery are often depicted as synonyms, which, from a terminological point of view, as mentioned earlier, raises the dilemma of whether apples are indeed being compared to apples. An analysis of the terms shows that the notion of a single access to the body (“single incision”, “single access”, “single port” or “single site”) seems to play a central role in this type of surgery. Two of the terms, however, refer to “incisionless” and to “natural orifice” surgery, respectively, which indicates the inexistence of an external incision, hence opposing the notion that prevails in the remaining designations. Additional information is provided by most terms as regards the location of the incision (“umbilical” or the more specific “transumbilical”), the type of surgery (the more generic “minimally invasive surgery” or the more specific “endoscopic” and, going further down the hierarchy, “laparoscopic”), and the use of a given type of equipment (“video”, “conventional equipment-utilizing”). This seems to point towards the idea that not all these designations are in fact representing the same concept, but a more thorough analysis, which is not within

¹¹ The Laparoendoscopic Single-Site Surgery Consortium for Assessment and Research (LESSCAR), that published a consensus statement with the main conclusions of that meeting (Gill *et al.*, 2010). The Urologic NOTES Group also endorsed LESS surgery as the designation for single-port surgery (Box *et al.*, 2008).

¹² The medical specialty more actively devoted to the diagnosis and treatment of Endometriosis.

the scope of this article, would be necessary in order to confirm this hypothesis and further develop it.

In order to get a glimpse of the actual usage of these terms among the community of subject field experts, a search was conducted in MEDLINE/PubMed® with the full forms of the 22 collected designations and resorting to the following search expression: ("Term" [All Fields] AND gynecology [All Fields]) AND ("2010/01/01" [Date-Publication]: "2016/03/01" [Date-Publication]). The aim was to see which terms have been more widely used in scientific, peer-reviewed papers within Gynecology since 2010 (the date of publication of LESSCAR's White Paper). The results showed that "laparoendoscopic single-site surgery" has become the most commonly used designation (with 90 results), followed by "single-port access" (45), "single-port laparoscopy" (29), "single-incision laparoscopic surgery" (27), and "single-port laparoscopic surgery" and "single-port surgery" (23 results each).

From the 90 scientific articles for "laparoendoscopic single-site surgery", only 15 had the full text freely available and were in English, so they constituted the selected corpus. The AntConc¹³ corpus analysis tool allowed the study of a set of definitions of "laparoendoscopic single-site surgery" and the subsequent extraction of data¹⁴, leading to the following lexical network:

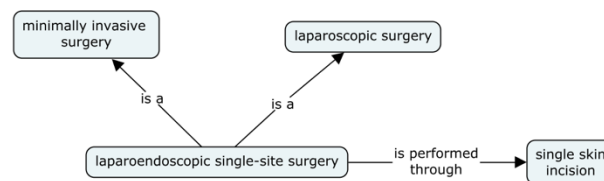


Fig. 1. Lexical network created with data extracted from the corpus

In this particular example, it seems that resorting only to a corpus-based approach, though useful, might be insufficient to fully grasp the notion of LESS surgery. Thus, there was a search for additional information in a set of biomedical terminological resources. Some of the resources were entirely hierarchical (such as ICD-10 and MeSH), others contained non-hierarchical relations as well (e.g. SNOMED CT and UMLS). Current procedure classifications were also consulted, namely the NOMESCO Classification of Surgical Procedures (2012); the German Procedure Classification (Operationen- und Prozedurenschlüssel – OPS, 2016 version); the OPCS Classification of Interventions and Procedures, version 4, used by the UK's National Health Service; and the French Classification Commune des Actes Médicaux (2016 update). Being a type of surgery that is estimated to account for 50-80% of current surgeries in some medical specialties (particularly urology and gynecology) (cf. Gill *et al.*, 2010), it was surprising to realize that the concept as such does not exist at the moment in any of the

¹³ Available at <http://www.laurenceanthony.net/software.html>.

¹⁴ *Is a* and *is performed through* were actually present in the corpus and hence express lexical relations. The former should therefore not be confused with the conceptual *is_a* relation.

consulted resources. Although data on <Laparoscopy> or <Minimally invasive surgery> are available, there is nothing that refers specifically to a single incision, which would allow concept differentiation. The inclusion of additional content about <laparoendoscopic single-site surgery> in one of these resources was believed to be pertinent, and SNOMED CT has been chosen based on the arguments presented earlier.

Bearing all of this in mind, supplementary searches were conducted, showing the need to go beyond the verbal and incorporate non-verbal (images or diagrams) as well as multimodal elements¹⁵ about LESS surgery (within the context of Gynecology). The data analysis led to the creation of a set of concept maps using the OTE tool¹⁶, which were then validated by two senior expert gynecologists. Due to space limitations, only one of the maps will be shown. The map below (Fig. 2) depicts the concept under analysis and aims to position it within the broader concept of <Surgical procedure> by making use of a specific differentiation, Aristotelian-based approach. It confirms that the lexical network from Fig. 1, although incomplete, contains elements that may indeed correspond to relevant characteristics of the concept under analysis. As a matter of fact, the existence of a single skin incision constitutes the essential characteristic (cf. ISO 1087-1: 2000) of this type of surgical procedure.

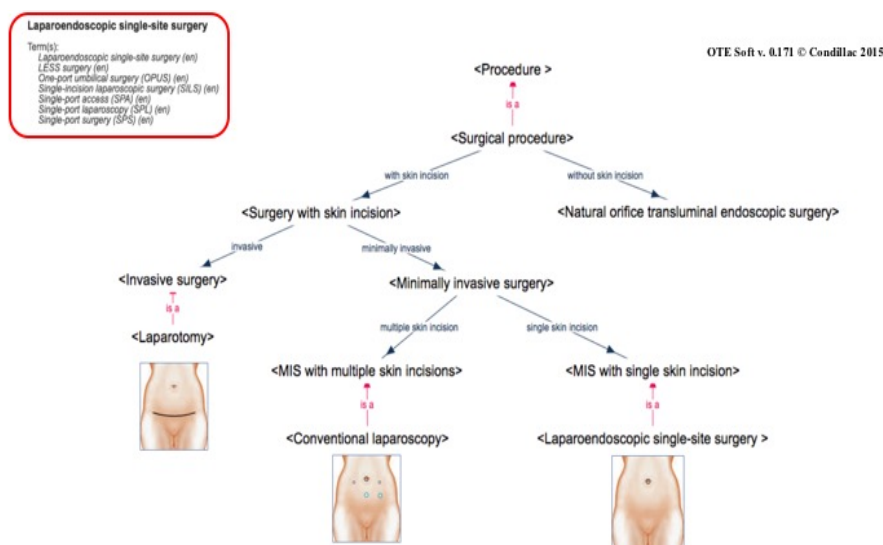


Fig. 2. Concept map of <Laparoendoscopic single-site surgery>

Taking into account both the linguistic and conceptual information gathered thus far, and after analyzing SNOMED CT's categorial structure for procedure concepts, it is

¹⁵ Namely medical video articles, a new type of scholarly communication that has been more thoroughly described in Carvalho, Roche, and Costa (forthcoming).

¹⁶ Created by the Condillac research group, from Université de Savoie Mont Blanc. The maps, as well as the tool, have been described in more detail in Carvalho, Roche, and Costa (2015).

believed that this resource could benefit from the inclusion of the concept <Laparoendoscopic single-site surgery>, given the increasing prevalence of this type of surgery in some medical specialties and the likely need to refer specifically to LESS surgery in EHRs. Therefore, our proposal would be as follows: as regards the descriptions¹⁷ (the linguistic dimension), the choice of the Fully Specified Name and respective Synonyms would respect the position issued by LESSCAR and supported by our MEDLINE/PubMed® searches, as seen below (Fig. 3).

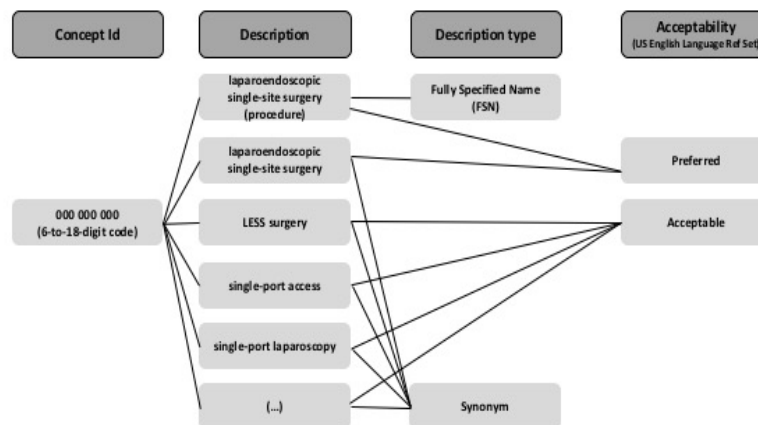


Fig. 3. Example of descriptions for <Laparoendoscopic single-site surgery>

As for the conceptual dimension, which is the basis of concept definitions in SNOMED CT, the biggest challenge lies in the current absence, in this resource, of any attribute-value relationship referring to the essential characteristic of the concept in question, i.e. the **single skin incision**. Although the concept of <Incision of skin (procedure)> exists, its subtype concepts are related to the location of the incision, with no data concerning the number of incisions. The same happens with <Incision – action (qualifier value)>, which refers to the method used in practically all surgical procedures. Since <Incision of skin (procedure)> would not work as a valid destination concept for the attribute relationships used to define procedure concepts, as it would conflict with the domain constraints, it is believed that the concept of <**Single incision – action (qualifier value)**> should be introduced in SNOMED CT, in order to enable concept differentiation and, hence, concept definition.

The concept of <Laparoendoscopic single-site surgery> could therefore be defined through a combination of is_a and attribute relationships, represented in both a human and computer-readable way via SNOMED CT's compositional grammar, which supports interoperability¹⁸. The following proposal (Fig. 4) resorts, as much as possible, to

¹⁷ For consistency purposes, SNOMED CT's terminology will be maintained in our proposal. Therefore, instead of adopting the notion of designation (cf. ISO 1087-1), the original expression "description" will be used.

¹⁸ For further information, cf. IHTSDO (2015).

currently existing concepts, descriptions and relationships. The suggestions have been signaled in red.

```

000000000 |Laparoendoscopic single-site surgery (procedure)|

=== 264274002 |Endoscopic operation (procedure)| +
51316009 |Laparoscopic procedure (procedure)| +
7113634006| Surgery using robotic assistance (procedure)| :
{ 260686004 |Method (attribute)| = 000000000 |Single incision – action (qualifier value)| },
405813007 |Procedure site - Direct (attribute)| = 113345001 |Abdominal structure (body structure)| +
12921003 |Pelvic structure (body structure)| +
51185008 |Thoracic structure (body structure)|,
424226004 |Using device (attribute)| = 86174004 |Laparoscope, device (physical object)| +
82830000 |Robotic arm, device (physical object)|,
42876005 |Surgical approach (attribute)| = 103388001 |Percutaneous approach – access (qualifier value)|

```

Fig. 4. <Laparoendoscopic single-site surgery> using SNOMED CT's compositional grammar

The first three concepts represent the *is_a* relationships (types of surgery) and are followed by a refinement, which is introduced by a colon and consists of a sequence of one or more attribute-value pairs. The attribute is separated of the value by an equals sign and if there is more than one value for the same attribute, the plus sign is added. The different attribute-value pairs are separated by commas. Curly braces represent grouping of attributes within a refinement, for example to indicate that a given method applies to a specific site.

5 Concluding remarks

By resorting to a case study, this paper aimed to reflect upon the fact that analyzing the conceptualization of a given subject field and the corresponding discourses produced by the expert community may result in representations that do not always match, but both play a vital role in terminology work: through ontologies, conceptualization proposals open new possibilities in terms of interoperability by resorting to the Semantic Web and W3C standards; albeit with vagueness and inconsistencies, the discourses provide fundamental access to the expert community as a way to stabilize knowledge in different areas of expertise, which is particularly relevant in new techniques or approaches, as is the case of LESS surgery. When anchored in this double dimension, terminology work may contribute to further enhance that stability and, consequently, the quality of specialized communication.

Acknowledgements

This research has been financed by Portuguese National Funding through the FCT Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade Nova de Lisboa – UID/LIN/03213/2013.

References

- Autorino, R. *et al.* (2011). “LESS: An Acronym Searching for a Home.” *European Urology* 60 (6): 1202–4.
- Box, G. *et al.* (2008). “Nomenclature of Natural Orifice Transluminal Endoscopic Surgery (NOTES) and Laparoendoscopic Single-Site Surgery (LESS) Procedures in Urology.” *Journal of Endourology* 22 (11): 2575–81.
- Carvalho, S., C. Roche, and R. Costa (2015). “Ontologies for Terminological Purposes: The EndoTerm Project.” In *Proceedings of the 11th International Conference on Terminology and Artificial Intelligence - Universidad de Granada, Granada, Spain, November 4-6, 2015*, edited by T. Poibeau and P. Faber: 17–27. Granada: CEUR Workshop Proceedings.
- . (forthcoming). “Why Read When You Can Watch? Video Articles and Knowledge Representation within the Medical Domain.” In *Proceedings of the 2015 TOTH Conference*. Chambéry: Équipe Condillac / Université Savoie-Montblanc.
- Cimino, J. (1998). “Desiderata for Controlled Medical Vocabularies in the Twenty-First Century.” *Methods of Information in Medicine* 37 (4-5): 394–403.
- . (2001). “Terminology Tools: State of the Art and Practical Lessons.” *Methods of Information in Medicine* 40 (4): 298–306.
- Coiera, E. (2015). *Guide to Health Informatics*. 3rd edition. New York: CRC Press.
- Commission Nationale d’Évaluation des Dispositifs Médicaux et des Technologies de Santé. “Classification Commune Des Actes Médicaux (CCAM).” Accessed February 25, 2016. <http://www.ameli.fr/accueil-de-la-ccam/index.php>.
- Costa, R. (2013). “Terminology and Specialised Lexicography: Two Complementary Domains.” *Lexicographica* 29 (1): 29–42.
- Deutsches Institut für Medizinische Dokumentation und Information. “Operationen- Und Prozedurenschlüssel (OPS).” Accessed March 3, 2016. <https://www.dimdi.de/static/de/klassi/ops/kodesuche/onlinefassungen/opshtml2016/index.htm>.
- Duclos, C. *et al.* (2014). “Medical Vocabulary, Terminological Resources and Information Coding in the Health Domain.” In *Medical Informatics, E-Health: Fundamentals and Applications*, edited by A. Venot, A. Burgun, and C. Quantin: 11–41. Paris: Springer.
- Escobar, P. and T. Falcone, eds. (2014). *Atlas of Single-Port, Laparoscopic, and Robotic Surgery - A Practical Approach in Gynecology*. New York: Springer.
- Georgiou, A. *et al.* (2012). “Evolution and Simplified Terminology of Natural Orifice Transluminal Endoscopic Surgery (NOTES), Laparoendoscopic Single-Site Surgery (LESS), and Mini-Laparoscopy (ML).” *World Journal of Urology* 30 (5): 573–80.
- Gill, I. *et al.* (2010). “Consensus Statement of the Consortium for Laparoendoscopic Single-Site Surgery.” *Surgical Endoscopy* 24 (4): 762–68.

- International Health Terminology Standards Development Organisation (2014). "SNOMED CT Starter Guide." Accessed May 12, 2015. http://ihtsdo.org/fileadmin/user_upload/doc.
- (2015). "SNOMED CT Compositional Grammar Specification and Guide." Accessed March 4, 2016. http://ihtsdo.org/fileadmin/user_upload/doc.
- (2016). "SNOMED CT Browser." Accessed February 25, 2016. <http://browser.ihtsdotools.org/>.
- International Organization for Standardization (2000). "ISO 1087-1: Terminology Work - Vocabulary - Part 1: Theory and Application." Geneva: International Organization for Standardization.
- (2007). "ISO 17115: Health Informatics - Vocabulary for Terminological Systems." Geneva: International Organization for Standardization.
- National Health Service (2016). "OPCS-4 Classification of Interventions and Procedures." Accessed March 3, 2016. http://www.datadictionary.nhs.uk/web_site_content/supporting_information/clinical_coding/opes_classification_of_interventions_and_procedures.asp.
- NIH and US National Library of Medicine (2016). "MeSH Browser (2016)." Accessed February 25, 2016. <https://www.nlm.nih.gov/mesh/MBrowser.html>.
- Ramesh, B., M. Vidyashankar, and P. Dimri (2014). *Single-Port Laparoscopic Surgery in Gynecology*. New Delhi: Jaypee Brothers Medical Publishers Ltd.
- Roche, C. *et al.* (2009). "Ontoterminology: A New Paradigm for Terminology." In *International Conference on Knowledge Engineering and Ontology Development, Oct 2009*: 321–26. Funchal.
- Roche, C. (2012). "Ontoterminology: How to Unify Terminology and Ontology into a Single Paradigm." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, May 23-25, 2012*, edited by N. Calzolari *et al.*: 2626–30. Istanbul: European Language Resources Association (ELRA).
- (2015). "Ontological Definition." In *Handbook of Terminology - Vol. 1*, edited by H. J. Kockaert and F. Steurs: 128–52. Amsterdam: John Benjamins Publishing Company.
- Santos, C. and R. Costa (2015). "Domain Specificity: Semasiological and Onomasiological Knowledge Representation." In *Handbook of Terminology - Vol. 1*, edited by H. J. Kockaert and F. Steurs: 153–79. Amsterdam: John Benjamins Publishing Company.
- Springborg, H., and A. Fader (2015). "Laparoendoscopic Single-Site Surgery: LESS, General Indications." In *Minimally Invasive Gynecological Surgery*, edited by O. Istre: 157–62. Heidelberg: Springer.
- World Health Organization (2016). "ICD-10 Version:2016." Accessed February 25, 2016. <http://apps.who.int/classifications/icd10/browse/2016/en>.
- WHO-FIC Collaborating Centre in the Nordic Countries (2012). "NOMESCO Classification of Surgical Procedures." Accessed March 4, 2016. http://www.nordclass.se/ncsp_e.htm.

Describing Knowledge Organization Systems in BARTOC and JSKOS

Andreas Ledl¹ and Jakob Voß²

¹ University Library of Basel, Basel

² Verbundzentrale des GBV (VZG), Göttingen

Abstract. This paper introduces a cooperation between the Basel Register of Thesauri, Ontologies & Classifications (BARTOC) and project coli-conc to provide information about Knowledge Organization Systems, which “encompass all types of schemes for organizing information and promoting knowledge management” (Hodge 2000), in uniform form. The result is a proper metadata scheme, the JSKOS data format, and an API to connect and access connecting terminology registries so terminologies can be discovered and explored at one place.

Keywords: knowledge organization systems · terminology registries · metadata schemes

1 Introduction

Over the last twenty-five years a large amount of Knowledge Organization Systems (KOS) such as classifications, thesauri, authority files, and term bases have been published online and new ones are added almost daily. Several terminology registries have emerged to identify, describe and make accessible these KOS, ideally in a human- and machine-readable way. These registries replaced link lists, which usually contained information about only a few well-known controlled vocabularies without elaborated search interfaces or bibliographic description of KOS. The BARTOC terminology registry³ has quickly evolved to one of the largest collections of information about distinct KOS. This paper summarizes the description of KOS in BARTOC and project coli-conc,⁴ and provision of its metadata as Linked Open Data and the uniform JSKOS data format.

3. <http://bartoc.org/>

4. <https://coli-conc.gbv.de/>

2 The Basel Register of Thesauri, Ontologies & Classifications (BARTOC)

According to Golub et al., who identified four types of KOS registries (Metadata Registries, basic or full Terminology Registries, Service Registries and Data Registries), BARTOC is a basic Terminology Registry, because it contains “only the metadata of KOS vocabularies” (Golub et al. 2014). Furthermore, it is a meta registry of KOS registries (see figure 2), linking to 70 other portals.⁵ BARTOC differs from other terminology registries on five counts: it includes *any kind* of KOS from *any subject area* in *any language*, *any publication format*, and *any form of accessibility*. This means that it needs universal systems for formal cataloging, classification and subject indexing of knowledge organization systems.

2.1 The origins of BARTOC

The idea for BARTOC has its roots in two classic areas of Library & Information Science: creating bibliographies and teaching information literacy. On the one hand, it is the latest contemporary descendant of intensive efforts in the 20th century to publish printed surveys of the work on KOS. On the other hand, controlled vocabularies are needed to tag pieces of information and to apply complex search strategies like the “block building approach”, where a topic is broken down into separate sections to analyze the scope (termino)logically.

It was clear from the start that BARTOC would address the international library community, but also terminologists and scientists from all over the world. Since its launch in November 2013, it has had a total of 500'000 visits and 3.3 million page views.

2.2 BARTOC's current metadata scheme

BARTOC contains “a relatively sufficient amount of metadata” (Bratková and Kučerová 2014). The metadata scheme used to describe KOS in BARTOC originates from the early days when BARTOC was just a blog called “Thesaurusportal”.⁶ With migration of the database

5. <http://bartoc.org/en/terminology-registries>

6. <http://www.profi-wissen.de/hilfsmittel-fuer-alle-denkbaren-recherchegebiete-thesaurus-porta/>

to Drupal CMS the schema was extended with a mapping to RDF, so KOS description in BARTOC can be used as Linked Open Data. Table 1 lists all current metadata fields including their mapping to JSKOS (see section 3 and figure 3 later) and RDF. The mapping to RDF makes use of schema.org, FOAF, and SKOS ontology.

Table 1. Metadata schema and mappings of KOS description in BARTOC

Field	JSKOS	RDF
URI	<code>uri</code>	subject URI
Title	<code>prefLabel</code>	<code>skos:preflabel</code> , <code>schema:name</code>
Alternative or English Title	<code>altLabel</code>	<code>skos:altLabel</code> , <code>schema:name</code> , <code>dct:title</code> , <code>foaf:name</code>
Author	<code>creator</code>	<code>dct:creator</code> , <code>schema:creator</code>
Abstract	<code>scopeNote</code>	<code>skos:scopeNote</code> , <code>dct:description</code>
Coverage	<code>subject</code>	<code>dct:subject</code>
Type	<code>type</code>	<code>dct:type</code> , <code>rdf:type</code>
Format	-	<code>dct:format</code>
Size	<code>extent</code>	<code>dct:extent</code>
License	-	<code>dct:license</code> , <code>schema:license</code>
Access	-	<code>dct:rights</code>
DDC	<code>subject</code>	<code>dct:subject</code> , <code>schema:about</code>
DDC Main Class	- ⁷	
Wikidata	<code>identifier</code>	<code>skos:exactMapping</code> , <code>dct:identifier</code>
Link	<code>url</code>	<code>schema:url</code> , <code>foaf:page</code>
Language	<code>language</code>	<code>schema:inLanguage</code> , <code>dct:language</code>
Topic	<code>subject</code>	<code>dct:subject</code> , <code>schema:about</code>
Year of Creation	<code>created</code>	<code>dct:created</code>
Term Translations	- ⁷	
VIAF	- ⁸	
Address	- ⁸	
Location	- ⁸	

2.3 Alignment with NKOS AP

Both BARTOC and JSKOS origin in a bottom-up process by actual description of knowledge organization systems. For this reason the current state is not finished until it has been tested sufficiently in several real-world applications. The Networked Knowledge Organization Systems Dublin Core Application Profile (NKOS AP), created between 2010 and 2015 followed the opposite direction by theoretical investigation of KOS and their registries. The resulting metadata

⁷. Only used for searching.

⁸. Not referring to the KOS but to its publisher.

scheme is expected to be “very important to terminology registries, service registries, vocabulary users (machine or human), and retrieval systems” (Zeng and Žumer 2015). A comparison of the current meta-data scheme of BARTOC, JSKOS, and NKOS AP resulted in an overlap at 13 of 28 fields for BARTOC and 18 for JSKOS (table 2).

Table 2. Mapping of NKOS AP to BARTOC and JSKOS

NKOS AP field	BARTOC	JSKOS
dct:title	Title	prefLabel, altLabel
dct:creator	Author	creator
dct:publisher	Author	publisher
dct:description	Abstract	scopeNote
dct:subject	Coverage, Topic, DDC	subject
dct:type	Type	type
dct:language	Language	languages
dct:identifier	URL, Wikidata	uri, identifier
dcat:contactPoint	Link	url
dct:license	License	license
nkos:sizeNote	Size	extent
dct:format	Format	-
dct:created	Year of Creation	created
dct:issued	-	issued
dct:modified	-	modified
wdrs:describedBy	-	subjectOf
dct:isPartOf	-	partOf
prov:wasDerivedFrom	-	versionOf
nkos:serviceOffered	-	concepts, types
dct:audience	-	not defined yet
nkos:basedOn	-	not defined yet
nkos:updateFrequency	-	to be discussed
nkos:usedBy	-	to be discussed
nkos:alignedWith	-	to be discussed
frbrer:isRealizationOf	-	to be discussed
frbrer:isEmbodimentOf	-	to be discussed
dct:relation	-	to be discussed
adms:sample	-	to be discussed

2.4 Use of controlled vocabularies to describe KOS

One particularly special feature of BARTOC, compared to other terminology registries, is its use of controlled vocabularies to describe KOS. It is considered as BARTOC’s “advantage that it specializes in supplementing Dewey’s decimal classification terms (up to the third hierarchic level) . . . , as well as providing the multilingual EUROVOC

thesaurus descriptors” (Bratková and Kučerová 2014). The KOS used to describe other KOS in BARTOC are described below. Each of them also has a BARTOC record in, given with its corresponding URI.

EuroVoc (<http://bartoc.org/en/node/15>) was chosen, although developed especially for the European parliamentary activities, because it is maintained by a trusted authority, it is open data, its domains are multidisciplinary and its terms are available in 25 languages, which is essential for BARTOC’s multilingual search. EuroVoc subject headings can be selected as *Topic* in Advanced Search.

DDC (<http://bartoc.org/en/node/241>) is the most widely used library classification system, translated in more than 30 languages. DDC codes up to the third hierarchy level enable grouping different KOS according to a certain field or topic. To make the search interface more easily accessible to wide-ranging groups of users, BARTOC provides DDC numbers and/or captions for content statistics, in the Advanced Search and in faceted search. The service is based on a subscription model. DDC was further expressed as Linked Data (Panzer 2013) and project coli-conc investigates the connection of DDC to other classification systems so it can be used as mapping backbone with other systems and content.

KOS Types Vocabulary (<http://bartoc.org/en/node/1665>) was developed by the DCMI NKOS Task Group (Dublin Core Metadata Initiative. NKOS Task Group 2015) and is, as far as we see, the only controlled vocabulary for KOS types. It differentiates between 14 different types of KOS (categorization scheme, classification scheme, dictionary, gazetteer, glossary, list, name authority list, ontology, semantic network, subject heading scheme, synonym ring, taxonomy, terminology, and thesaurus).

Wikidata (<http://bartoc.org/en/node/1940>) is a general purpose database and authority file that anyone can edit. By now BARTOC only contains mappings to corresponding KOS records in Wikidata to provide links to Wikipedia articles.

Additional vocabularies are used for format, license, and languages of KOS but they have not been published as terminologies yet.

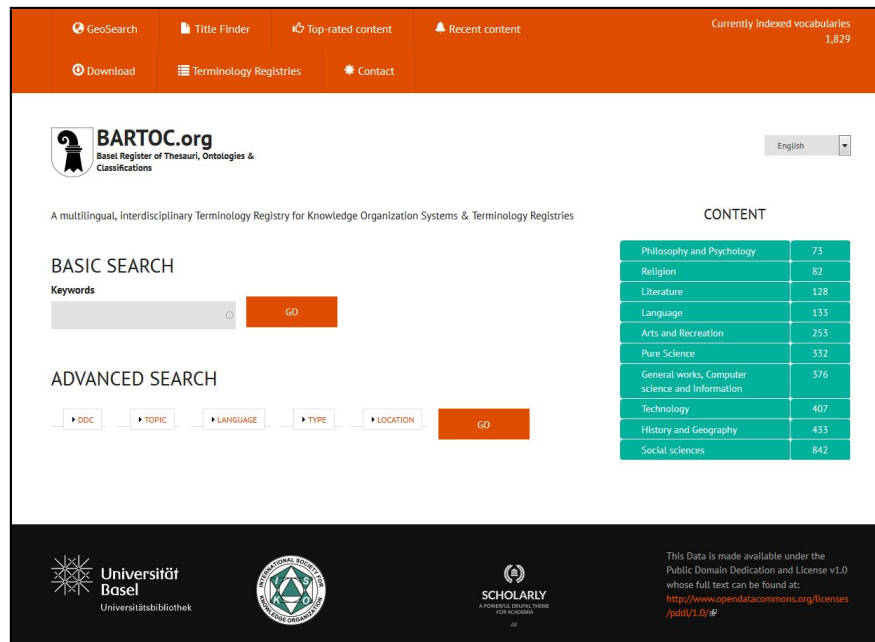


Fig. 1. Screenshot of BARTOC's search interface

3 JSKOS data format for Knowledge Organization Systems

The coli-conc project at Verbundzentrale des GBV (VZG) is funded by German Research Foundation (DFG) to facilitate management and exchange of concordances between knowledge organization systems. This includes the collection and provision of information about KOS and its concepts in a uniform format. To some degree such format is given with the Simple Knowledge Organization System (SKOS) ontology. SKOS allows the exchange of KOS as Linked Data on the Web but it comes with the complexity of RDF and it requires extensions to cover more than basic properties. To better support use of KOS data, especially in web applications, the JSKOS data format

for Knowledge Organization Systems is precisely defined, tested, and documented (Voß 2016c). JSKOS is also compatible with JSON-LD so it can be mapped to and from SKOS/RDF, if needed.⁹

3.1 JSKOS metadata scheme

In a nutshell, JSKOS supports the following object types:

- **Concepts** as basic entities of all KOS are covered well by SKOS. JSKOS only adds general fields from Dublin Core and common fields found in authority records.
- **Concept Schemes** are equivalent to KOS. In addition to descriptive fields a link to an API can be provided for querying concepts from this concept scheme. Figure 3 shows EuroVoc as example of a concept scheme expressed in JSKOS.
- **Concept Types** can be used to broadly group concepts, for instance concepts about places, people, events, and abstract topics.
- **Mappings** and **Concordances** describe mappings between concepts or concept schemes. This is a major contribution of JSKOS because support of mappings in plain SKOS is very limited.
- **Registries** collect concepts, concept schemes, concept types, mappings, concordances and/or other registries. Registries have no counterpart in SKOS neither.

Figure 2 illustrates the application of JSKOS objects to BARTOC. The website contains both a terminology registry and a meta registry of other terminology registries. Each KOS in BARTOC can be described as JSKOS Concept Scheme. The concepts of each KOS are not included in BARTOC but project coli-conc provides converters and mappings to make them accessible via downloads and an API. The metadata fields to describe objects in JSKOS are consistent for all object types, for instance `prefLabel` is used for both concept labels and concept scheme titles (see figure 3).

3.2 JSKOS-API

Reusing terminologies does not only require a uniform data format but also methods to access and query selected parts of a KOS. Such

9. JSON-LD defines general mapping rules from JSON to RDF. General JSON-LD, however, has too many degrees of freedom, in contrast to JSKOS.

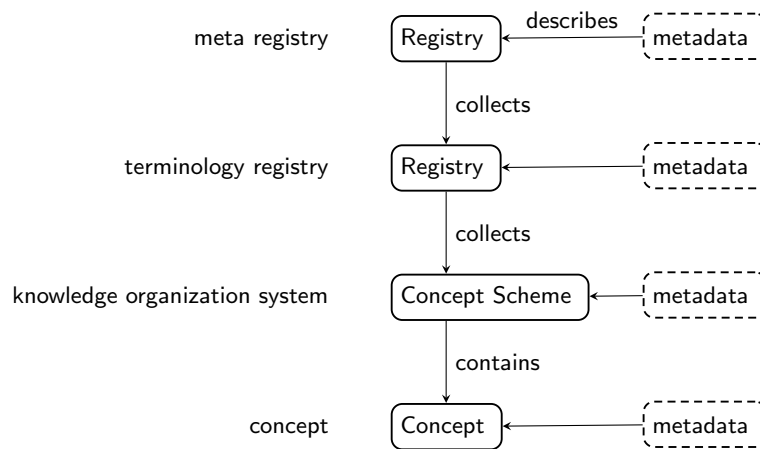


Fig. 2. Overview of metadata about KOS and registries

```

{
  "@context": "https://gbv.github.io/jskos/context.json",
  "id": "http://bartoc.org/en/node/15",
  "prefLabel": {
    "en": "Multilingual Thesaurus of the European Union"
  },
  "altLabel": { "en": "EuroVoc" },
  "url": "http://eurovoc.europa.eu/",
  "identifier": [ "http://www.wikidata.org/entity/Q1370467" ],
  "type": [
    "http://www.w3.org/2004/02/skos/core#ConceptScheme",
    "http://bartoc.org/en/taxonomy/term/1",
    "http://bartoc.org/en/taxonomy/term/2"
  ],
  "subject": [ {
    "id": "http://dewey.info/class/001",
    "prefLabel": { "en": "Knowledge" }
  }, {
    "id": "http://eurovoc.europa.eu/4060",
    "prefLabel": { "en": "European Union" }
  } ],
  "languages": [ "bg", "ca", "hr", "cs", "da", "nl", "en", "et", "fi",
    "fr", "de", "el", "hu", "it", "lv", "lt", "mk", "mt", "pl", "pt",
    "ro", "sr", "sk", "sl", "es", "sv" ]
}

```

Fig. 3. Abbreviated JSKOS record of Eurovoc terminology

methods can be provided either by downloading and importing the whole KOS into a database or by querying an existing service via API. Several APIs and services exist for selected KOS (for instance WebDewey¹⁰ for DDC) but without common standard and many terminology provider avoid the technical effort of setting up and maintain an additional web service. For this reason project coli-conc defines JSKOS-API based JSKOS and evaluation of similar APIs.

The full specification of JSKOS-API requires an ongoing overview of uses cases for terminology services (Voß 2016a). A subset of the most common requirements has already be defined as Entity Lookup Microservice API (ELMA) (Voß 2016b). The API provides two basic methods of access:

- **Entity Search** queries a list of concepts matching a query string with relevance ranking. The access method is intended for typeahead to select a concept of unknown URI. The response format is the same as OpenSearch Suggestions API (Clinton 2006).
- **Entity Lookup** queries one concept by its URI. The access method is intended to get details about a known concept.

JSKOS-API/ELMA services have been implemented as database application¹¹ and as wrappers¹² to access GND, Wikidata, ORCID, DDC and other KOS. The implementations are published as open source to be used in other applications as well.¹³

Based on JSKOS-API applications can make use of any KOS that is available in JSKOS format. As BARTOC is also mapped to JSKOS, it can be accessed by the same method. Planned applications at VZG include a tool to create and evaluate concept mappings, and a general terminology service (“Normdatendienst”) to provide a uniform search and browsing interface to multiple terminologies.

4 Summary

The Basel Register of Thesauri, Ontologies & Classifications prepares thousands of Knowledge Organization Systems under one interface in

10. <http://dewey.org/webdewey/>. This service is based on a subscription model.

11. See <https://github.com/gbv/cocoda-db>

12. See <https://jskos-php-examples.herokuapp.com/>.

13. See <https://coli-conc.gbv.de/publications/> for a current list of software.

order to achieve greater visibility, to highlight their features, to make them searchable and comparable, and to foster knowledge sharing.¹⁴ BARTOC covers a lot of user tasks, allowing “to find, identify, select, obtain ... KOS resources through the data provided” (Golub et al. 2014). When a user has found an interesting terminology, he or she is directed to the publisher’s site for further investigation. But once the KOS is made available via JSKOS-API, its concepts and structure can directly be explored from other places as well. The publication of more and more KOS via JSKOS-API, as being implement in project coli-conc, will allow users to directly browse and search in KOS from BARTOC. In reverse, the content of BARTOC registry will be searchable from other sites as well.

Due to the mutual benefit for both, BARTOC and coli-conc, it will be the most urgent task to improve alignment of BARTOC metadata scheme, NKOS AP metadata scheme, and JSKOS data format. The advantages of the latter, compared to plain RDF, include ease of use, a uniform description also for mappings, concordances, and registries, and a defined method to query registries and concept schemes. This way both BARTOC and JSKOS(-API) will foster the visibility, availability and usefulness of Knowledge Organization Systems in general.

References

- Bratková, Eva, and Helena Kučerová. 2014. “Knowledge Organization Systems and Their Typology.” *Revue of Librarianship* 25 (2): 1–25.
- Clinton, DeWitt. 2006. *OpenSearch Suggestions extension*. Technical report. <http://www.opensearch.org/Specifications/OpenSearch/Extensions/Suggestions/1.0>.
- Dublin Core Metadata Initiative. NKOS Task Group. 2015. *KOS Types Vocabulary*, October. http://wiki.dublincore.org/index.php/NKOS%5C_Vocabularies.

14. See (Hlava 2011)

- Golub, Koraljka, Douglas Tudhope, Marcia Lei Zeng, and Maja Žumer. 2014. "Terminology registries for knowledge organization systems: Functionality, use, and attributes." *Journal of the Association for Information Science and Technology* 65 (9): 1901–1916. doi:10.1002/asi.23090.
- Hlava, Marjorie. 2011. "Developing an Eclectic Terminology Registry." *Bulletin of the American Society for Information Science and Technology* 37 (4): 19–22.
- Hodge, Gail. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Washington, DC: The Digital Library Federation. <http://www.clir.org/pubs/reports/reports/pub91/contents.html>.
- Panzer, Michael. 2013. "DEWEY: how to make it work for you." *Knowledge Quest* 42 (2): 22–29.
- Voß, Jakob. 2016a. *Anforderungen an Normdatendienste*. Technical report 4. coli-conc, April. doi:10.5281/zenodo.50180. <https://dx.doi.org/10.5281/zenodo.50180>.
- . 2016b. *Entity Lookup Microservice API (ELMA)*. Technical report Version 0.0.3. March. <http://gbv.github.io/elma/>.
- . 2016c. *JSKOS data format for Knowledge Organization Systems*. Technical report Version 0.1.0. February. <http://gbv.github.io/jskos/>.
- Zeng, Marcia Lei, and Maja Žumer. 2015. *Networked Knowledge Organization Systems Dublin Core Application Profile (NKOS AP)*. Accessed January 28, 2016. <http://nkos.slis.kent.edu/nkos-ap.html>.

Toward Dynamic Representations of ThirdPlaceLearning

Mara Alagic, Tatiana Orel, & Glyn Rimmington

mara.alagic@wichita.edu; Wichita State University, USA
tatiana.orel@gmail.com; Algonquin College Language Institute, Canada
glyn.rimmington@wichita.edu; Wichita State University, USA

Abstract. Previously, Cognitive Linguistic methods were used to visually represent the ThirdPlaceLearning (TPL) theory of intercultural communication as a TPL general frame of terminology and concepts that are static. This visualization of the TPL theory comprised a set of mental/conceptual constructs, which were negotiated across the disciplines of intercultural communication and cognitive linguistics. In this report we introduce the steps leading to development of dynamic, entelechial representation of the TPL terminology system which can be used to yield new insights into application of frames, terminology management, ontologies and visualization of conceptual models.

Keywords: ThirdPlaceLearning (TPL), TPL frame, TPL ontology, TPL Relational Criteria

1 ThirdPlaceLearning Terminology System

The ThirdPlaceLearning (TPL) concept (Rimmington and Alagic 2008) represents a third, learned point of view that is distinct from the perspectives of oneself and the other. This learning comprises a proactive, Perspective Sharing and Perspective Taking (PSPT) strategy (Rimmington and Alagic 2008; Alagic, Rimmington and Orel 2009), which is facilitated by a set of processes and conditions known as the TPL relational criteria. These include active listening, dialectic flow of thinking, intercultural sensitivity, critical co-reflection, conscientization, and bodymindfulness (Alagic 2009). The novelty and complexity of the InterCultural Communication (ICC) learning makes it an ideal candidate for terminological analysis.

Ontologically, TPL terminology is a knowledge representation system comprising relevant concepts or entities as a network of semantic relationships between categories and frames. The term frame was first used by Minsky in 1975 as a paradigm to understand visual reasoning and natural language processing (Minsky 1977). In lexical semantics (Fillmore 1977), a frame is an abstraction of a form of mental representation encompassing one's experience,

knowledge and perception. Further, frames, as complex conceptual structures, are used to, represent categories for animates, objects, locations, physical events, or mental events (Barsalou 1992). The structural organization of frames is a network of nodes and their relations where nodes contain specific instances of data and capture an overarching, holistic representation of the TPL knowledge base (Rimmington and Alagic 2008).

The ThirdPlaceLearning structured terminology system derives from both a formal level (a thesaurus) and a cognitive/semantic level (concepts, categories and frames). The precise structure arises from categorization or clustering of like concepts. The TPL terminology system was represented as a structured domain of knowledge using cognitive linguistics methods (Alagic, Rimmington and Orel 2009; Orel, Alagic, and Rimmington 2014) and may serve as a basis of visualization of TPL concepts.

As a knowledge organization system, the TPL terminology system can be visualized as a complex network of conceptual components for understanding the TPL theory and its implementation in the intercultural learning context. In this form, the TPL terminology system is a constructivist learning tool that can support learners' preparation for intercultural communication. This form of terminological representation can improve conceptual clarity by characterizing the TPL domain's ontology in terms of generic concepts, their definitions and relationships (Benitez, Pilar and Prieto 2009).

The conceptual structure of the TPL terminology system is monolingual, non-hierarchical and explicitly defined terminologically. It is visually presented in Tables 1 and 2 as frames and subframes, which capture a range of conceptual relations—generic-specific and part-whole—which explicate the semantic and syntactic behaviors of these specialized language units.

The TPL theory was described formally (Table 1) with 18 terms that are part of the category *Abstract Concepts*. This category comprises two levels: basic (11 abstract concepts) and subordinate. The subordinate level comprises seven subcategories, based on functionality or type, namely:

- i. Mode of communication (e.g., *cultural dialectics*);
- ii. Miscommunication (e.g., *disorienting dilemma*);
- iii. Cultural/social influence (e.g., *conformity*);
- iv. Cultural representation (e.g., *iceberg*);
- v. Self (e.g., *self-identity*);
- vi. Proposition (e.g., *thesis*); and
- vii. Whole/part (e.g., *ThirdPlaceLearning*).

The latter subcategory, Whole/Part, has two terms: *ThirdPlaceLearning* and *TPL Relational Criteria*, which play a significant role in facilitating TPL processes, conditions and outcomes (Orel, Alagic, and Rimmington 2014).

Term	Category	Proposition
ThirdPlaceLearning	Abstract Concepts	[CONCEPT - BE ABOUT - WHOLE/PART]
Relational Criterion		[CONCEPT - BE ABOUT - WHOLE/PART]
Disorienting Dilemma		[CONCEPT - BE OF - TYPE/KIND]
Misconception		[CONCEPT - BE - WRONG]
Miscommunication		[CONCEPT - BE - WRONG]
Preconception		[CONCEPT - BE - WRONG]
Iceberg effect		[CONCEPT1 - CHANGE - CONCEPT2]
Self-Identity		[CONCEPT - BE ABOUT - SELF]
Active Listening	Process	[PROCESS - BE OF - TYPE/KIND]
Critical Coreflection		[PROCESS - BE OF - TYPE/KIND]
Perspective Sharing		[PROCESS - DEAL WITH - STATE/ABSTRACT CONCEPT]
Perspective Taking		[PROCESS - DEAL WITH - STATE/ABSTRACT CONCEPT]
Perspective Shift		[PROCESS - DEAL WITH - STATE/ABSTRACT CONCEPT]
Intercultural Sensitivity	State	[STATE (relation) - BE OF - TYPE/KIND]
Conscientization		[STATE1 - (NOT) BE AWARE OF - STATE2]
Bodymindfulness	Hybrid	[PROCESS/STATE - BE OF - TYPE/KIND]
Dialectic Thinking		[PROCESS/CONCEPT - BE OF - TYPE/KIND]
ICC Context		[TIME/SPACE/RELATION - BE OF - TYPE/KIND]
Multiple Perspectives		[STATE/CONCEPT - BE OF - TYPE/KIND]
Point of View		[STATE/CONCEPT - BE OF - TYPE/KIND]
Perspective		STATE/CONCEPT
Liminal Phase		[TIME/SPACE - BE PART OF - WHOLE]
Discourse (of learning)	Space	[SPACE - BE LIMITED BY - BOUNDARY]
Cognitive Learning Domain		[SPACE - BE OF - TYPE/KIND]
Psychomotor Learning D.		[SPACE - BE OF - TYPE/KIND]
Affective Learning Domain		[SPACE - BE OF - TYPE/KIND]
Interpersonal Learning D.		[SPACE - BE OF - TYPE/KIND]

Table 1. TPL frame building blocks: Categorized terms and related onomasiological models (propositions) (Orel, Alagic, and Rimmington 2014)

2 TPL System: General Frame

ThirdPlaceLearning (ABSTRACT CONCEPT₁) is ICC context-dependent, where *ICC context* encompasses spatial, temporal, relational, and historical sub-contexts (TIME/SPACE₂/RELATION). The *Discourse of TPL* (SPACE₁) encompasses the four *learning domains*—cognitive, psychomotor, interpersonal, and affective—(SPACE₂₋₅).

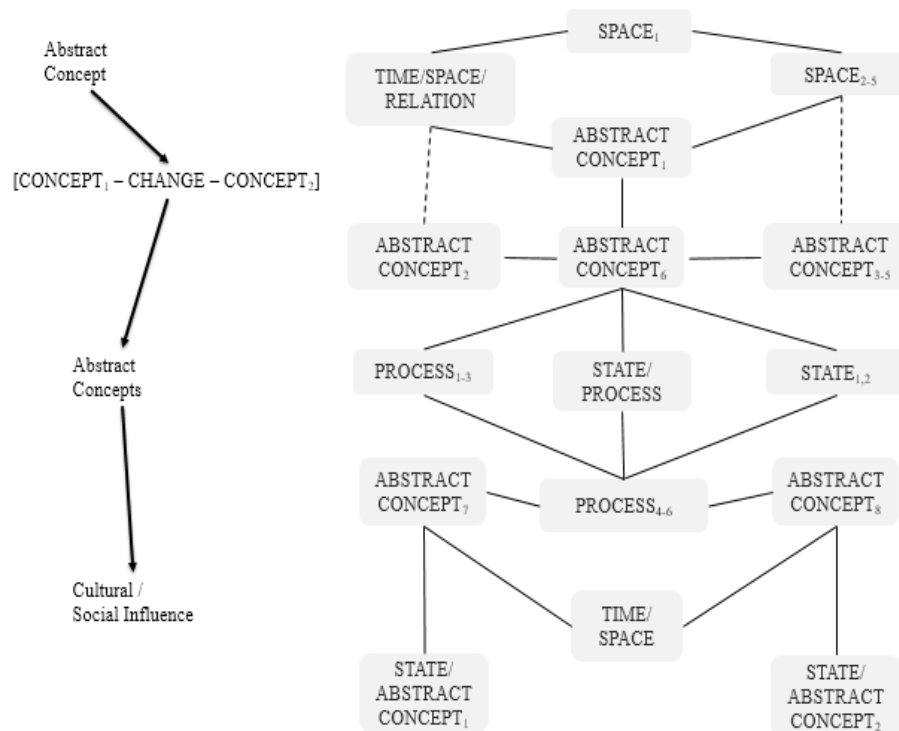


Fig. 1. TPL Frame Visualization – Conceptual Level
<http://tinyurl.com/hzvjpwl>

The transformational process of TPL may be triggered by *disorienting dilemma*, *misconception*, *miscommunication*, *preconception* (ABSTRACT CONCEPT_{3, 4, 5, 9}). The *iceberg effect* (ABSTRACT CONCEPT₂) metaphor captures the iceberg's deep, (subconscious) levels or *meaning structure*, which is a cultural lens that affects how we see other cultures.

During intercultural communication, the *liminal phase* (TIME & SPACE) for TPL transformation of an individual's *self-identity* (ABSTRACT CONCEPT₇) encompasses communicative and cognitive PROCESSES_{4, 5, 6} (*perspective taking*, *perspective sharing*, *perspective shift*). Transformation of self-identity involves changes in *perspective*, *point of view*, or *multiple perspectives* (STATE & ABSTRACT CONCEPT₁). The above processes are facilitated by the TPL relational criteria.

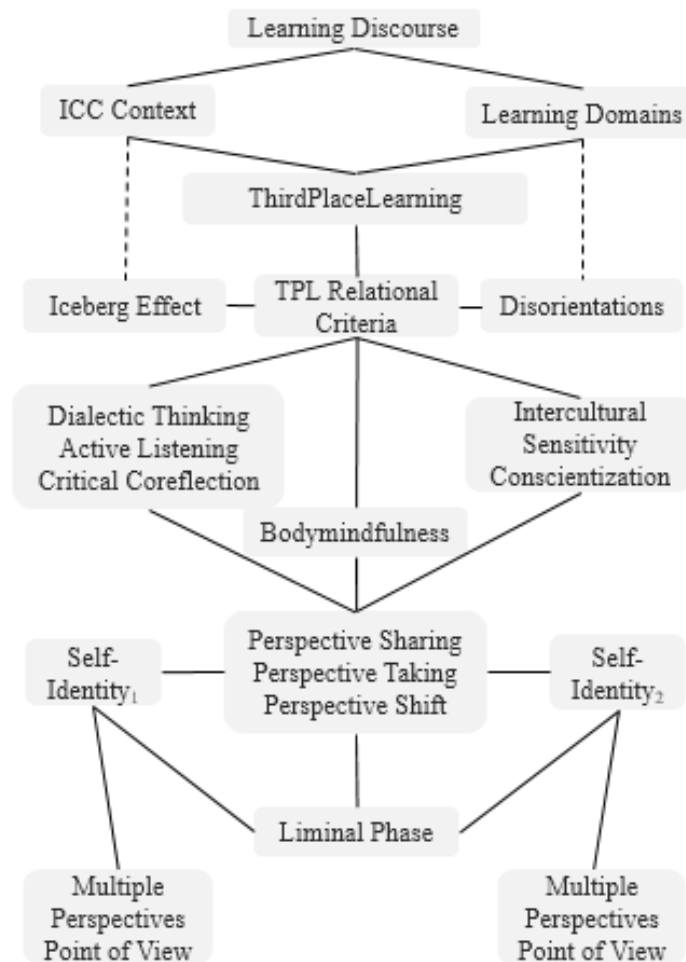


Fig. 2. TPL Frame Visualization –Terminological Level <http://tinyurl.com/hzvjpwl>

TPL relational criteria (ABSTRACT CONCEPT₆) comprise: cognitive and communicative PROCESSES_{1, 2, 3} (*dialectic thinking, active listening; critical co-reflection*); awareness and somatic-emotional STATES_{1, 2} (*intercultural sensitivity, conscientization*); and a conceptual hybrid STATE & PROCESS of the somatic-emotional process and awareness state (*bodymindfulness*).

The TPL frame (Figures 1 & 2) is a mental representation of the interconnected concepts of Perspective Sharing and Perspective Taking (PSPT)

strategy (Alagic, Rimmington and Orel 2009) and six enabling relational criteria (subframes) that lead to a third point of view during intercultural interactions.

2.1 Relational Criteria Example: Intercultural Sensitivity

Term	Category	Proposition
Intercultural Sensitivity	State	[STATE – BE OF – TYPE/KIND]
Respect	State	
Tolerance	State	
Oblivious	State	[STATE1 – (NOT) BE AWARE OF – STATE2]
Ambiguity	Property	[PROPERTY – BELONG TO – INFORMATION/ COMMUNICATION]
Avoiding vs. Seeking	Process	[PROCESS1 – OPPOSE TO – PROCESS2]
Iceberg	Abstract Concept	
Cultural mélange		[CONCEPT – BE OF – TYPE/KIND]
Meaning Structure	Hybrid	[STATE/CONCEPT – BE USED FOR – OPERATION]
Aware & Invalidate	Hybrid	[STATE – BE WITH – ACTION]
Aware & Dismiss	Hybrid	[STATE – BE WITH – ACTION]
Learn & Validate	Action	[ACTION1 – BE WITH – ACTION2]
Learn & Celebrate	Action	[ACTION1 – BE WITH – ACTION2]

Table 2. Intercultural Sensitivity subframe building blocks: Categorized ICS terms and related onomasiological models (propositions) (Orel, Alagic, and Rimmington 2014)

The *Intercultural Sensitivity* (Alagic, Orel, and Rimmington 2010) subframe (of the TPL frame) represents an associative network of dialogically negotiated mental/conceptual constructs and is shown in Table 2 as a list of defined terms, categories and propositions. The categories arose from clustering of similar underlying concepts.

To better understand intercultural sensitivity as defined in Table 2, it is useful to consider the point of view of an AGENT, which is grounded in his/her meaning structure, cultural iceberg and cultural mélange. The *Meaning structure* (STATE & CONCEPT) is a collection of assumptions, perspectives, and expectations that act as a filter for the AGENT's interpretation of the world and originate during the AGENT's developmental years. The AGENT's cultural *Iceberg* (ABSTRACT CONCEPT₂) is the metaphoric representation of culture that captures visible and invisible layers, the accessible and less-accessible cultural/identity beliefs and values. Finally, the AGENT's *cultural mélange*

(ABSTRACT CONCEPT₃) is a metaphor for an AGENT's complex mixture of cultural experiences.

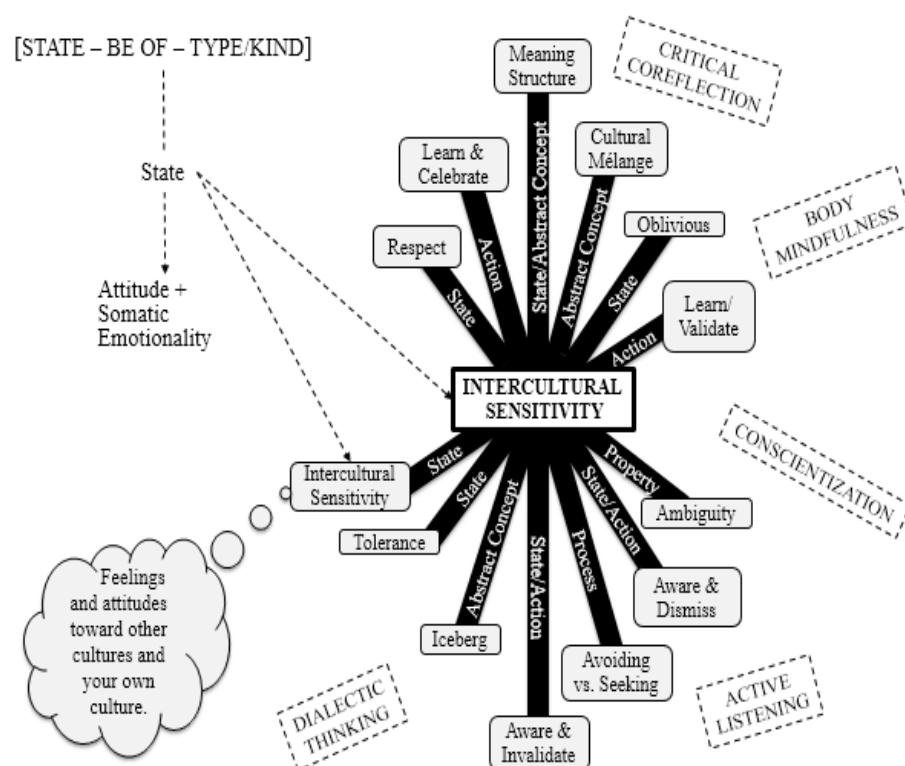


Fig. 3. Intercultural Sensitivity subframe within TPL relational criteria subframes at the terminological level; Connection to the TPL frame at the conceptual level is also indicated via the appropriate proposition STATE-BE OF-TYPE/KIND.

Figure 3 represents the intercultural sensitivity subframe at the terminological level, including line labels for categories and corresponding subframe terms.

Dynamic representations of the TPL frame (Figure 2) and the six TPL relational criteria subframes, including for Intercultural Sensitivity (Figure 3) are available at <http://tinyurl.com/hzvjpwl> for learners to explore TPL in more detail at a linguistic/terminological and ICC level.

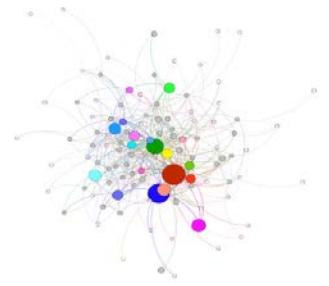
3 Conclusions

In this report we introduced the steps leading to development of dynamic, entelechial representation of the TPL terminology system. A frame of the TPL theory comprised six relational criteria subframes. We explored the InterCultural Sensitivity (ICS) subframe in particular. First the TPL frame and ICS subframe were presented as terms, their categories and corresponding onomasiological models (propositions) (Tables 1 and 2). Then these frames were presented visually (Figures 1, 2 & 3). Further, these are available in a dynamic form online (<http://tinyurl.com/hzvjpwl>). Together, these provided new insights into application of frames, terminology management, ontologies and visualization of conceptual models.

REFERENCES

1. Alagic, Mara. 2009 (November 21–24). Third Place Learning: Relational Criteria for Developing Intercultural Capital. Paper presented at the *Academy of Management Annual Meeting - Professional Development Workshop: Teaching and Learning in Different Cultures*, Chicago, Illinois.
2. Alagic, Mara, Orel, Tatiana, and Rimmington, Glyn M. 2010. Frames as a Model for Understanding Cultural Sensitivity (Abstract). *Proceedings of 7th International Conference on Intercultural Communication Competence*, 198-207. Khabarovsk, Russia, September 14-16.
3. Alagic, Mara, Rimmington, Glyn M., and Orel, Tatiana. 2009. Third Place learning Environments: Perspective Sharing and Perspective Taking. *International Journal of Advanced Corporate Learning*. 4 (2). Accessed April 15, 2016, <http://online-journals.org/i-jac/article/view/985>.
4. Barsalou, Lawrence W. 1992. Frames, Concepts, and Conceptual Fields. In *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, edited by Eva Feder Kittay and Adrienne Lehrer, 21-74. Hillsdale, NJ: Lawrence Erlbaum Associates.
5. Benitez, Pamela F., Pilar, León-Araúz, and Prieto, Juan A. 2009. Semantic Relations, Dynamicity, and Terminological Knowledge Bases. *Current Issues in Language Studies* 1. Accessed April 17, 2016, <http://www.academicpress.us/journals/511X/download/v1n1-1.pdf>

6. Fillmore, Charles J. 1977. Topics in Lexical Semantics. In *Current Issues in Linguistic Theory*, edited by R. Cole, 76-138. Bloomington, IN: Indiana University Press.
7. Minsky, Marvin. 1977. Frame Theory. In *Thinking: Readings in Cognitive Science*, edited by Philip N. Johnson-Laird and Peter C. Wason, 355-376. Cambridge: Cambridge Press.
8. Orel, Tatiana, Alagic, Mara, and Rimmington, Glyn M. 2014. Concept System Analysis of the ThirdPlaceLearning Theory. *Terminology and Knowledge Engineering* 2014. Germany: Berlin. Accessed April 19, 2016, <https://hal.archives-ouvertes.fr/hal-01005870>
9. Rimmington, Glyn and Alagic, Mara. 2008. *Third Place Learning: Reflective Inquiry into Intercultural and Global Case Painting*. In book series Huber-Warring, T. Teaching <~> Learning Indigenous, Intercultural Worldviews International Perspectives on Social Justice and Human Rights. Charlotte, NC: Information Age Publishing Inc.



Making the Visualization of Concepts More Attractive and Smarter

Concepts are said to be abstract and general. This makes them so useful but at the same time difficult to grasp. Therefore there have always been attempts to render concepts more “intuitive” and easier to understand by providing them with a visual representation instead of definitions and other textual explanations. Examples abound: ontologies, the diagrams of, e.g., satellite systems for concept analysis (Nuopponen 2010), Euler and Venn diagrams (cf., e.g., Hammer 1995, Moktefi and Shin 2013), the existential graphs of Ch. S. Peirce (Roberts 1973, Queiroz and Stjernfelt 2011) as well as conceptual graphs in logic (cf., e.g., Sowa 1984), the drawings of elementary geometry (Miller 2007), the Hasse diagrams of lattice theory and formal concept analysis (Ganter and Wille 1999), and the diagrams of category theory in mathematics, the structural formulas of chemistry, the force diagram of Lewin’s vector psychology, the network graphs used in both computer science and sociology (as well as in other disciplines). The illustrations in Wüster’s (1968) machine tool dictionary are another good example of visualization and nonverbal representation of concepts.

Drawing techniques, however, are just one type of visualization techniques, others include photography, film and animation. Their importance for human cognition and communication has recently attracted a renewed and more intensive attention from different areas of education, business and research as testified, for instance, by recent proposals for visual representation in terminology. This trend, which has been labeled by such terms as “the pictorial turn”, “iconic turn”, or “visual turn” and which has given rise to the transdisciplinary endeavour of “visual culture studies”, mirrors the increasingly significant role played by visuals in today’s digital society. Via the ubiquitous World Wide Web, images can be distributed globally and using a wide range of digital media and platforms we can access and view these images in a number of ways and in a number of different situations.

The purpose of this workshop is to attract experts from a variety of research areas to participate in an interdisciplinary effort to share and discuss how to make more attractive and smarter visualization techniques that can, in turn, significantly help to represent and communicate more effectively information in different domains of knowledge.

Organizing Committee

Professor Klaus Robering (robering@sdu.dk), University of Southern Denmark

Associate Professor Lotte Weilgaard Christensen, (lotte@sdu.dk), University of Southern Denmark

Associate Professor Rocio Chongtay, (rocio@sdu.dk), University of Southern Denmark

Professor Bettina Berendt, (Bettina.Berendt@cs.kuleuven.be), KU Leuven, Belgium

References

1. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Berlin and New York (1999)
2. Hammer, E. M.: Logic and Visual Information. Stanford: CSLI (1995)
3. Miller, N.: Euclid and his Twentieth Century Rivals. Diagrams in the Logic of Euclidean Geometry. Stanford: CSLI (2007)
4. Moktefi, A., Shin, S.-J. (eds.): Visual Reasoning with Diagrams. Birkhäuser, Basel (2013)
5. Queiroz, João; Stjernfelt Frederik (eds.): Diagrammatical Reasoning and Peircean Logic Representation. Issue 136 of *Semiotica* (2011)
6. Roberts, D. D.: The Existential Graphs of Charles S. Peirce. Mouton, Paris and The Hague (1973)
7. Sowa, J. F. (1984): Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, Reading, MA (1984)
8. Nuopponen, A.: Methods of concept analysis - towards systematic concept analysis. Part 2 of 3. The LSP Journal - Language for special purposes, professional communication, knowledge management and cognition, 1(1), 5–14 (2010) Available online: <http://rauli.cbs.dk/index.php/lspcog/index>
9. Wüster, E.: The Machine Tool: An Interlingual Dictionary of Basic Concepts. Technical Press, London (1968)

Satellite System as a Visualization Tool for Concept Analysis

Anita Nuopponen

University of Vaasa, Finland
Anita.nuopponen@uva.fi

Abstract. The paper deals with the so called *satellite model* which is a type of visual method for analyzing concepts with a special emphasis on concept relations and concept systems. It integrates terminological methods and principles with a visualization technique and is designed for systematic terminology work and concept analysis. The satellite model method draws on an existing classification of relation types while other mind mapping and concept mapping methods are based on user generated relation types. Initially, it was introduced with pen and paper in mind in the 1980's, but when combined with mind mapping software it becomes an even more flexible way to analyze and visualize concepts and various relation types between them in any field of knowledge. This makes it applicable for various other purposes, too (e.g. research, writing articles, teaching, planning, designing, coordination, system design, information design and modeling, technical communication).

Keywords: concept, concept system, satellite model, satellite system, terminology, concept mapping, mind mapping

1 Introduction

This paper discusses the *satellite model*, which is a type of visualization and analysis tool or method for terminological concept analysis developed in the 1980's and first described in [1]. The product of the process is a *satellite system*¹, which consists of a *core concept* and concepts in *satellite nodes*, which in turn may have their own satellite nodes etc. A satellite system can represent either a homogenous concept system with the same type of relations (e.g. generic or partitive concept system) between concepts, or a mixed concept system with various types of concepts which can be combined to the core concept and to each other with any type of relations (e.g. generic, partitive, causal, instrumental, temporal, relations) in a single diagram. [1,2]

The method reminds *mind mapping* [e.g. 3] and *concept mapping* [e.g. 4], but has been developed independently from them. While the others have their origins in educational purposes, the satellite model has its roots in terminology work and is based on terminological methods and principles. In this paper I will describe the background

¹ Often the presentation itself is called *satellite model*, too. In this paper, a distinction between the method (*satellite model*) and the result (*satellite system*) has been made.

and principles of the method and compare it with these other two visualization methods.

2 Background

The reason behind the creation of the satellite model was the inadequateness of the existing graphical presentation types for concept systems. There were no standardized diagrams for relations and concept systems other than generic and partitive ones. There did not exist any comprehensive visualization tool that would be able to combine separate but connected concept systems in order to create an overall picture of a delimited domain for a terminology project.

In terminological literature, various types of graphical presentations for concept system have been listed and exemplified [e.g. 5,6,7]. Traditionally, however, mostly only tree diagrams have been utilized in practice (see Fig. 1). They appear in two forms: diagrams with diagonal lines between nodes for generic (logical) concept relations (e.g. *diving suit* – *wetsuit*), and bracket diagrams with vertical and horizontal lines between the nodes for the partitive concept relations (e.g. *open-circuit* – *diving cylinder*; see also [8]. Both of these tree diagram types can be presented either horizontally or vertically, or combined as in Fig.1.

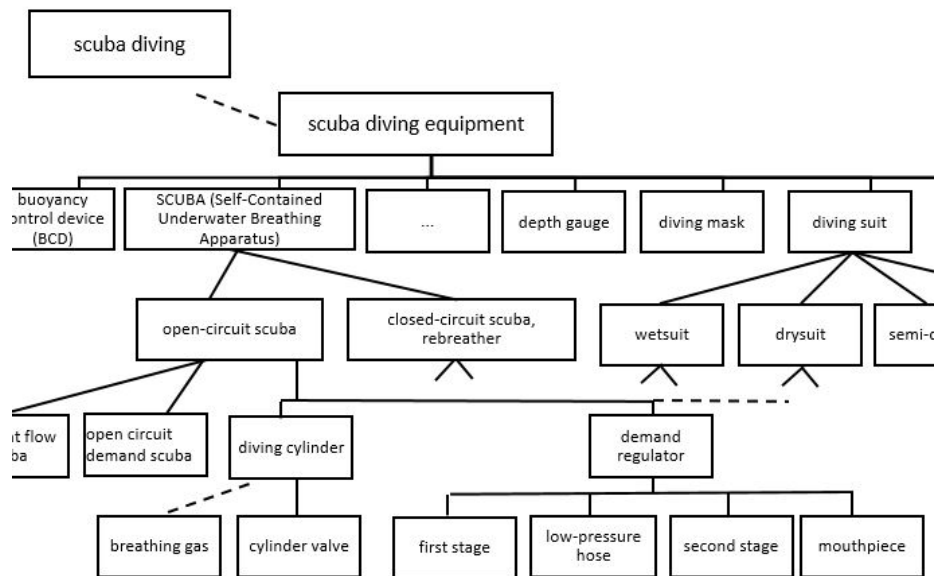


Fig. 1. A part of a mixed concept system: scuba diving equipment

The rest of the relation types are normally classified as associative relations, and marked with arrows or dotted lines in diagrams [e.g. 8] and connected to either a generic or partitive concept system (e.g. *scuba diving* – *scuba diving equipment*). Alternative presentations for time-related concept systems (e.g. temporal, developmental

and causal) could be flowcharts [see 7]. Various types of field diagrams, charts, (thesaurus-style) lists and tables are possibilities mentioned in the literature. However, there have not been any established ways for visualizing various types of associative concept relations and systems.

It is a challenge to try to combine all the central concepts of the selected field in the same concept diagram such as in the Fig. 1. This can be seen clearly in the case of projects where a large amount of candidate terms has been (automatically) extracted from a corpus and listed in alphabetical order, after which it may be an overwhelming task to (re)construct a concept system for further analysis and definition writing.

The satellite model was conceived originally for the pen and paper method during our terminology project courses. Together with the students we started to organize their separate concept system tree diagrams on a large paper sheet so that the shared superordinate concept was inserted in the middle of the paper and the tree diagrams were arranged in a circle around it (e.g. *road* and its various typologies according to different subdivision criteria). During the last three decades the model has been established as a tool for concept analysis and conceptual research utilized on courses in terminology and communication studies at the University of Vaasa.

3 Satellite as a Visual Metaphor

Visual metaphors are defined by Eppler and Burkhard [9] as “graphic depictions of seemingly unrelated graphic shapes (from other than the discussed domain area) that are used to convey an abstract idea by relating it to a concrete phenomenon”. The satellite system receives its name and form from another visual metaphor than the traditional presentation forms for concept systems. Traditionally the classic tree structure has been the basic visualization tool for terminological concept analysis. The tree is normally presented upside down with roots cut off. However, this type of tree presentation was experienced as too rigid to accommodate various types of concept relations. The satellite metaphor brought more flexibility and was adopted as the basis for the designation.

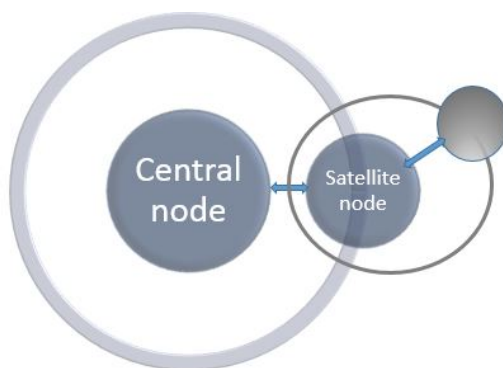


Fig. 2. The basic idea of the satellite model

The basic idea of the satellite system or model goes back to the presentations of a smaller object, i.e. a *satellite*, (e.g. the Sun) orbiting a bigger object (e.g. the Earth) (Fig. 2). Each satellite may have their own *satellites* (e.g. the Moon). The satellites are kept in their orbit by *gravity*. When transferring this model into terminology work, the Sun is the main concept (*core concept*) and its satellite is a concept that is connected to it by the “force” of the conceptual relationship between them.

Concept relations can be seen as the gravitational force holding the concepts together. Further than this it is needless to go into the metaphor, because not all the facts about the natural, or man-made satellites are comparable with concepts and their relations to other concepts. It was mainly the image of satellites orbiting the Earth that gave the idea of the satellite model: the closest concepts are in a closer orbit around the core concept in the *central node* and more distant concepts in an orbit further away, or travelling around a *satellite node*. In section 5, these nodes will be dealt in greater detail.

4 Building a Satellite System

4.1 Core Concepts and Satellite Concepts

In a satellite system, a single concept is taken as the core concept in the same way as in mind mapping [3] and concept mapping [see e.g. 10]. However, the purpose is different in these methods. In concept mapping as presented by Novak, the aim is to develop or test the student’s knowledge. The core concept gives the starting point together with a question to be answered by using the mapping process. [10] In using the satellite model as a method, the selection of the core concept serves to delimit the domain to be analyzed. Especially in the first phase of terminological concept analysis, the core concept of the satellite system represents the whole domain to be explored. It may be on a high level of abstraction or otherwise central to the field in question and above all it has to be able to link together the concepts and concept systems to be covered in the presentation. Depending on the concept analysis needs at hand, the core may refer e.g. to a discipline or another area of expertise as e.g. *scuba diving* in Fig. 3,4, or a part of these. It can represent concrete objects, activity, action, process, procedure (e.g. *scuba diving certification*), state, property etc. The core concept may thus refer either to a material object (*scuba diving equipment*) or an immaterial object (e.g. *scuba skills*). [See also: 2]

In the satellite system, one concept is taken in focus at a time as the core (*scuba diving* in Fig. 3). The concepts in the main satellite nodes (*diver*, *dive*, etc.) correspond to the main elements (divers/diver types; dives/diver types etc.) of the reference object of the core concept. The terms or other concept designations in the nodes can be written in singular to refer to the concept level (*diver*) or in plural (divers) to refer to the object level. Especially in the beginning of the analysis it is more natural to use plural in accordance with the source texts (e.g. “scuba divers can be either professional or recreational”) instead of formal expressions of the relations between concepts (e.g. “the subordinate concepts of *scuba diver* are *professional scuba diver* and *recreational scuba diver*”). In terminology work, the singular form is preferred in the final

version, because the terms in glossaries and term base entries are in singular (with some exceptions).

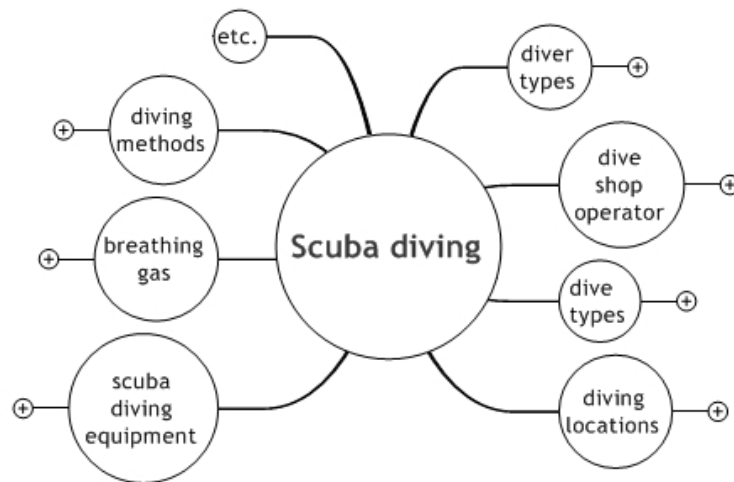


Fig. 3. The core concept and satellite nodes of a satellite system

The concepts in the satellite nodes get their own satellites, which again may get their own satellites, etc. In a satellite system, the relations (“gravitation”) can be of any type. The main satellite nodes together with the preliminary concepts orbiting them can be and ought to be separated from the main system as their own satellite systems for a more detailed analysis.

If the analysis combines two or more equally important departure concepts, it is recommended that they are dealt in their own satellite systems in order not to complicate the system too much. Sometimes, when the analysis proceeds and more information is compiled, another concept may attract more concepts than the core concept. Then the analyzer must consider changing the core concept or adjusting the focus.

4.2 Auxiliary Nodes

In terminological concept analysis it is necessary to distinguish between different types of concept relations and provide the diagrams with labels for division criteria etc. Because the purpose was to keep the satellite system visualization simple and flexible, no specific markers for each concept relation or system type were included in the method. Instead, auxiliary nodes with expressions for relations (e.g. *Who?*, [agent relation] in Fig. 4), division criteria (e.g. “according to the purpose”, “structure” etc.) and sometimes even characteristics are used (see Fig. 4). Also colored lines can be utilized to express relations but this requires a separate legend explaining the color code. These are solutions for including the concept relation information in the presen-

tation drawn by hand or by mind map software. When presenting satellite systems with e.g. XML or ontology software it is possible to have separate codes for different concept relation types.

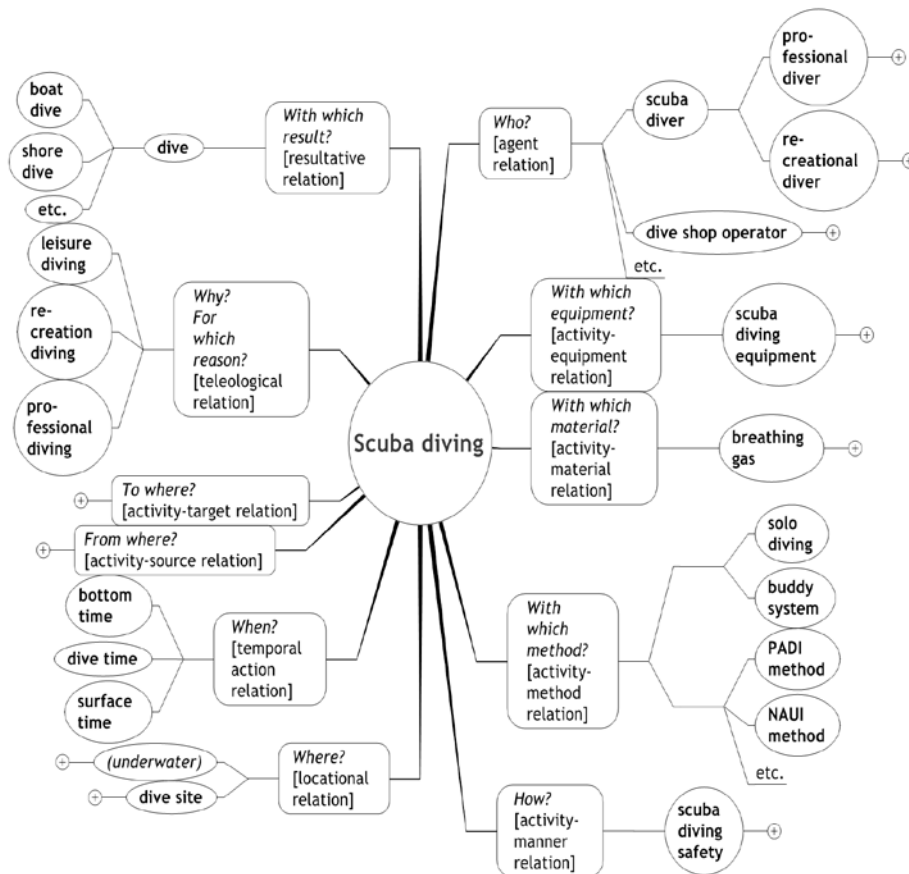


Fig. 4. A satellite system of scuba diving.²

In Fig. 4, the auxiliary nodes appear as the main satellites and contain questions concerning various types of concept relations (*With what? How?*) and the name of the relation (*tool relation; activity-equipment relation*). The concepts proper appear in the oval nodes in the figure (*scuba diving equipment*). The relations can also be expressed by concept roles, e.g. *agent, tool, result*. If the relation type is either obvious or diffi-

² A modified version of the figure in [16]. Further material for the examples has been taken from Wikipedia.

cult to define, it can be left out – especially in the beginning of the analysis when there is not yet enough information about all the concepts used. In Fig. 4, the generic relations have not been marked separately as such (e.g. *scuba diver* – *professional diver*). The emphasis is on the various points of views that can be taken when analyzing the concept of scuba diving. Concept relations and various types of starting points (typologies, structure, origination, activity, development, causation, transmission etc.) for concept analysis by using the satellite model approach are dealt more thoroughly in Nuopponen [2].

5 The satellite Model as a Tool for Terminology Work

Satellite model method can be used as a visualization tool almost for all the phases of terminology project and work in one way or another: e.g. as a mind mapping type of brainstorming tool [cf. 3] for planning and preliminary mapping of the domain as well as for delimiting the scope of the project in its initial stages, as a tool for extracting concepts, candidate terms, equivalents, and material for definitions etc., and for preliminary organizing manually or automatically extracted information. Later it can function as a tool for analyzing characteristics and delimiting concepts from each other, and for refining the concept system by specifying the relations between the concepts, as well as for presenting the end result, etc. [See 10]

In a terminology project, graphical concept system presentations are vital for project members so that they can understand each other and communicate more easily. [11]. In addition to projects covering a large numbers of concepts, satellite systems can be helpful for translators, terminologists and e.g. technical communicators or journalists doing ad hoc research for finding out translation equivalents for a smaller number of terms or definitions for a few concepts.

6 The Satellite Model and other Concept Mapping Methods

As mentioned above, the satellite model resembles many other visual methods utilized to map concepts and ideas. The best-known are mind mapping which was originally presented by Tony Buzan and concept mapping by Joseph D. Novak and his colleagues. In addition to them there are various other similar methods. Most of the mapping methods are developed for educational purposes and focus on teaching and learning specific subject matter and they are based on the assumption that “[if] students can represent or manipulate a complex set of relationships in a diagram, they are more likely to understand those relationships, remember them, and be able to analyze their component parts” as Davies [12] sums it up.

The satellite model was also born in an educational context but not for teaching or learning knowledge associated with subject matter but as a general concept analysis tool for terminology work and projects in any knowledge domain. Even though the satellite model has many similarities with mind mapping and concept mapping, it was the terminological methods and needs for visualization that were behind its development. It was the late 1980’s when we started to draw these models on our courses, and

at that time mind maps and their techniques were not yet familiar to us. Since then, maybe some 500–800 students have analyzed domains of their choice with satellite models for their coursework for Concept Analysis courses. Later on, we have utilized various types of mind mapping software (e.g. MindManager, Freemind) for drawing the satellite systems which has had some influence on the methods.

The methods of mind mapping and concept mapping differ in precision and formality. Concept maps are more formal and structured [12]. Mind maps have an organic structure and also otherwise emphasize visual means for remembering, e.g. pictures, thickness of the lines, colors etc. [3,12] (see Fig. 5). Just like the satellite method, the mind map method is based on the idea that the most important concept or theme is to be positioned in the middle. In the mind maps single word notations written on the lines are recommended and cross-referencing is welcomed. Eppler [13: 203] summarizes mind maps as “A mind map is a multi-coloured and image-centred, radial diagram that represents semantic or other connections between portions of learned material hierarchically.”

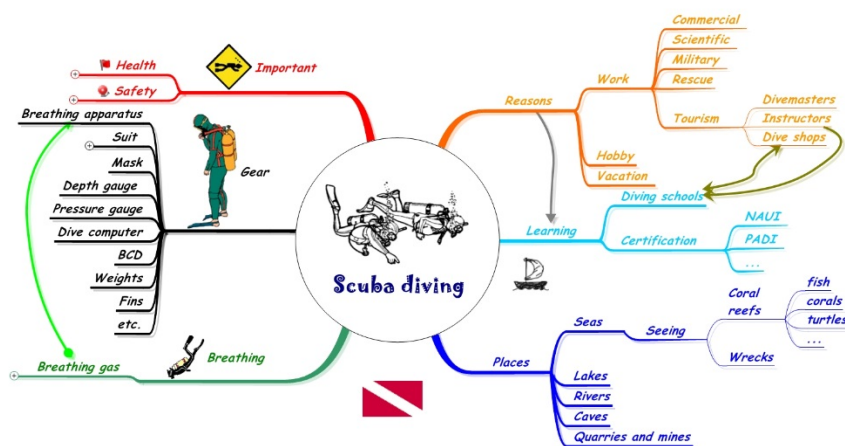


Fig. 5. Mind map of scuba diving

Eppler [13: 203] describes the concept map as a “top-down diagram showing the relationships between concepts, including cross connections among concepts, and their manifestations (examples)”. The concept maps utilize often the classic tree metaphor and organize concepts in a hierarchy where the most general concepts are on the top and those more specific are on a lower level below it [see 4: 1–2]. This resembles the traditional way of visualizing generic and partitive concept systems in terminological literature, and consequently shares the same space related problems. A concept map fragment is given in Fig. 6.

Cross-links are an important characteristic in concept maps. They link together “concepts in different segments or domains of the concept map”. [10: 2] In the satel-

lite systems, cross-linking is to be avoided in order to keep the presentation tidy and easy to read. Instead, one concept can be placed in several locations in the diagram. However, it can be defined only once, so it is good to mark the most relevant location for future stages of the project. In addition, it should be tested if the concept it is actually the same – or is it just the term, the linguistic designation, that is the same while the underlying concepts are different.

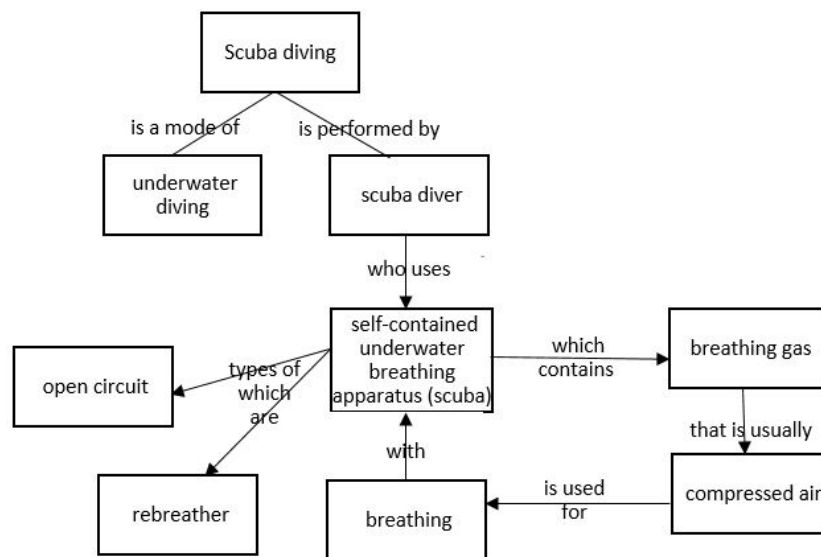


Fig. 6. Concept map: What is scuba diving?

In the concept maps, phrases are used for explaining the relations between concepts (e.g. *is comprised of, are involved in, are, need, require*) [4], which resembles in function the alternative relation markers in satellite systems described above. However, the satellite method is integrated as a tool into the systematic concept analysis method which provides the analyzer with a classification of various types of concept relations [2,10,14]. Prior knowledge of these helps the analyzer to quickly find the right place for a concept in the satellite system whereas a student analyzing a text and drawing a concept map, according to Novak and Cañas [4], has to find, define and name the relationship between the concepts. They have noticed that students “often comment that it is hard to add linking words onto the ‘lines’ of their concept map. This is because they poorly understand the relationship between the concepts, or the meanings of the concepts, and it is the linking words that specify this relationship. Once students begin to focus-in on good linking words, and on the identification of good cross-links, they can see that every concept could be related to every other concept.” [4: 13]

Also the satellite model can be utilized quite well for educational uses, even though the actual purpose of terminological concept analysis is to investigate concepts and terms of a field in order to present them in glossaries or term bases. The satellite model method provides analyzers with tools so that they do not need to create their own relationship vocabulary every time. The set of concept relations listed e.g. in [1,2] is extensive and covers most of the relation types, but it leaves also room for defining additional relation types and specifying the existing ones according to the authentic material.

7 Conclusion

Picht and Draskau [6: 64] list the following principles for concept system presentations which also apply for compiling satellite systems: (1) clarity: even a non-expert can get a quick and thorough idea of the special field; (2) intelligibility: user-friendly presentations avoid excessive complexity by limiting number of concepts and relations; (3) transparency: transparent and clearly understandable relation types and the classifying criteria; and finally (4) extendable without requiring overall revision. In addition, the satellite model is easy to learn to use, the most elaborate being to learn to distinguish between the different relation types [see 2], but it is not necessary always to utilize the whole set of relation types. The satellite model does not however require that one learns dedicated representations for every concept relation type and concept system type.

Compared to the traditional tree diagrams [e.g. in 8] a satellite system is flexible because new nodes can be added more easily starting from the middle instead of the top especially when drawing by hand on paper. The representation can be expanded, modified and specified during the analysis process. Nodes can be analyzed separately in their own satellite system and brought back together to form the whole picture of the field. In addition, any mind mapping software can be utilized and its additional features can enhance the usability of satellite systems. For instance, some mind mapping software offer the possibility to integrate a note with text and images in the nodes which enables the use of a satellite system as a tool for compiling information on concepts and terms, or even editing the final terminological product. Furthermore, the presentations can be exported in addition as a linear text document also as a set of interlinked web pages with the satellite system as the basic structure (see e.g. [15], which contains students' terminological vocabularies).

The satellite model has over the years proved to be a practical tool for visualizing concepts, concept systems, terminologies as well as special field knowledge. However, it can and must be developed further. Except for the other visual presentation tools discussed here there are today many interesting methods to be scrutinized and compared.

References

1. Nuopponen, A.: Begreppssystem för terminologisk analys. [Concept systems for terminological analysis] Acta Wasaensia. University of Vaasa, Vaasa (1994)
2. Nuopponen, A.: Methods of concept analysis - tools for systematic concept analysis (part 3 of 3). In: The LSP Journal - Language for special purposes, professional communication, knowledge management and cognition, 2011, 2(1): 4-15. <http://lsp.cbs.dk> (2011)
3. Buzan, T.: Mind mapping. <http://www.tonybuzan.com/about/mind-mapping/> (2011)
4. Novak, J. D., Cañas, A.J.: The Theory Underlying Concept Maps and How to Construct and Use Them. Technical Report IHMC CmapTools 2006-01 Rev 01-2008. <http://cmap.iuhmc.us/docs/pdf/TheoryUnderlyingConceptMaps.pdf> (2008)
5. Felber, H.: Terminology manual. Unesco, Paris (1984)
6. Picht, H., Draskau, J.: Terminology: An Introduction. University of Surrey, Surrey (1985)
7. Wright, S.E.: Representation of Concept Systems. In: Wright, S.E., Budin, G.: Handbook of Terminology Management, pp. 89–97. Benjamins, Amsterdam, Philadelphia (1997)
8. ISO 704:2009 Terminology work – Principles and methods. Geneva: ISO (2009)
9. Eppler, M. J., Burkhard, R.A.: Knowledge Visualization. Towards a new discipline and its fields of application. http://doc.rero.ch/record/5196/files/1_wpca0402.pdf (2004)
10. Nuopponen, A.: Methods of concept analysis - towards systematic concept analysis (part 2 of 3). In: The LSP Journal - Language for special purposes, professional communication, knowledge management and cognition, 2010, 1(1): 5-14. <http://lsp.cbs.dk> (2010)
11. Nykänen, O.: Sanastoprojektin vaiheet. In: Toimikunnista termitalkoisiin, pp. 62–71. The Finnish Centre for Technical Terminology, Helsinki (1999)
12. Davies, M.: Concept mapping, mind mapping and argument mapping: what are the differences and do they matter? In: Higher Education 62: 279–301 (2011)
13. Eppler, M.: A comparison between concept maps, mind maps, conceptual diagrams, and visual metaphors as complementary tools for knowledge construction and sharing. In: Information Visualization 5: 202-210 (2006)
14. Nuopponen, A.: Concept Relations v2. An update of a concept relation classification. In: Nistrup Madsen, B., Erdman Thomsen, H. (eds.) Terminology and Content Development, pp. 127–138. Association for Terminology and Knowledge Transfer (2005)
15. WasaTerm – A collection of students' terminology projects. University of Vaasa. <http://lipas.uvasa.fi/termino/WasaTerm/sanastot/> (2004–2016)
16. Nuopponen, A.: A model for structuring concept systems of activity. In: Wang, Y., Wang, Y., Tian, Y. (eds.) Terminology, Standardization and Technology Transfer, Encyclopedia of China Publishing House, Beijing (2006)

Simple Graphical Representations of Ontology-based Clinical Decision Support Knowledge Assets

Margarita Sordo¹²³, Christopher J. Vitale¹, Priyaranjan Tokachichu¹, Dan Bogaty¹,
Saverio M. Maviglia¹²³, Roberto A. Rocha¹²³

¹ Clinical Informatics, Partners eCare, Partners Healthcare, Boston, MA, USA;

² Brigham and Women's Hospital, Boston, MA, USA;

³ Harvard Medical School, Boston, MA, USA

{msordo, cjvitale, ptokachichu, dbogaty, smaviglia, raro-
cha}@partners.org

Abstract. We present simple graphical representations of different aspects of an ontology-based conceptual schema to represent clinical knowledge for decision support in the form of *if...then* production rules. We posit that a “simple is better” depiction, if well placed, not only may be more powerful in conveying the intended explanation, but more importantly, it may trigger the reader’s own mental processes and interpretations, resulting in a more complete understanding of the topic at hand. This approach facilitates understanding the complexity of the underlying model and its dependencies, and facilitates rule authoring. The ultimate goal is to foster consistency in rules implementation and maintenance; develop authoritative knowledge repositories to promote quality, safety and efficacy of healthcare; and enable future work in knowledge discovery.

Keywords: Visualization, knowledge representation, ontologies, clinical decision support, healthcare.

1 Introduction

As ontologies become more complex, the lack of simple visual displays hinder a clear understanding of any given model. Achieving a simple visualization is difficult and requires a higher understanding of the intricacies of the model before it is translated to a simple visual display for the end user. Simple visualizations go beyond a mere elimination or substitution of parts. Trade-offs are often required between what should be displayed (included) and what should be removed (eliminated) in any visualization, to ensure a complete and accurate understanding of the model being presented. The resulting simple visualization allows for a “whole view” of the model being communicated by assembling its different aspects [1].

Knowledge representation in healthcare is similarly intricate. The need for simple representations for modeling and curating complex clinical knowledge is essential for ensuring quality, safety and efficacy of healthcare. Knowledge assets are augmented with clinically-relevant concepts from platform-independent, standards-based models representing standard terminologies (e.g. LOINC, UCUM, SNOMED-CT, MeSH).

As we expand our models and link them to other ontologies, the complexity increases considerably. This is particularly true for Clinical Decision Support (CDS) production rules where both the antecedent and the consequent draw dependencies from multiple sources, e.g. electronic patient records, guidelines, information models, terminologies.

Users also may have different needs in terms of understanding ontology-based clinical content. For example, a clinician (knowledge viewer) may be more interested in viewing the clinical content of a CDS rule; whereas a knowledge engineer (knowledge author) may require a deeper understanding of the model and the components of a CDS rule.

The work reported herein is part of ongoing efforts by the Partners eCare Clinical Informatics Group at Partners HealthCare to represent all current CDS knowledge assets into the Clinical Knowledge Management System (CKMS) [2]. CKMS utilizes an ontology-based model that supports our continuing work on leveraging collective knowledge [3]. Such knowledge enables institutions across Partners HealthCare to effectively utilize and share knowledge-driven computer systems to promote continuous learning, overcome patient safety and quality challenges, and embrace new care delivery models and scientific advances.

The remainder of this paper is organized as follows: Section 2 describes our current work and presents a description of the proposed “simple is better” visualization approach. Section 3 presents a series of visual depictions of relevant sections underlying the production rule schema for CDS. Section 4 presents a visualization of an instance of a production rule, and in Section 5 we discuss the proposed approach.

2 Current Work

2.1 Visualization

As described in the previous section, our major visualization challenge relies on finding a simple, yet powerful representations capable of conveying the necessary information for users to piece together the individual (simpler) parts of a composite model. When necessary, these visualizations must couple visual depictions with descriptive narrative to further facilitate understanding and exploration.

Our proposed approach is based on the premise that “simple is better” visual representations. Such visual representations may facilitate understanding of complex ontologies by a) partitioning the model into a series of simple depictions; and b) abstracting away certain intricacies.

Other visualization efforts tend to overcrowd the display by presenting complete depictions of complex models [4][5][6]. Our strategy departs from these approaches in that we have chosen to visualize our models through simple representations of relevant components without overcrowding the display. Our belief is that these depictions will reduce the reader’s potential cognitive overload, resulting in a better understanding of an otherwise too complex representation of an underlying model [7].

2.1 Knowledge Modeling

One of the driving forces behind CKMS focuses on the formalization, or *conceptualization* of clinical knowledge assets, resulting in a “metamodel” that supports the definition of a set of conventions, elements, and types common to clinical data domains (e.g. patient demographics, medications, diagnoses). All models derived from the metamodel are declarative in nature, providing the building blocks for a flexible representation of knowledge assets across all domains. One such type of knowledge asset is the *production rule*. A production rule is a decision rule with an “*if*” part, or antecedent, and a “*then*” part, or consequent. The antecedent could be a simple expression or a Boolean combination of simple expressions, while the consequent represents one or more actions (e.g. notification, assertion, modification or retraction of facts), or some other side-effect. In other words, production rules are logic statements that specify the execution of one or more actions when their conditions are satisfied. The next section briefly describes the main elements of the current production rule schema (asset type), with particular emphasis on the antecedent of the rule.

3 Production Rule Schema

Based on the metamodel, the current schema for production rules consists of generic and specific properties. Generic properties contain information regarding provenance consistent with the Provenance Ontology proposed by W3C [8], and pre-defined constraints (e.g. clinical settings and patient-specific characteristics) for the overall rule or its discrete components. Specific properties model the rule expression, i.e. the data declarations, the logic expression or conditions in the antecedent of the rule, and the constraints where such conditions apply [9]; and actions and execution constraints [10] for the consequent of the rule. These specific properties define a new asset type and lay the foundation for modeling the logic of a rule and its behavior. Our approach is to separate each “component” and provide a simpler view of an otherwise overwhelming model. This also enables each component to evolve independently and be referenced (reused) by other knowledge assets. The remaining of this section explains these properties in detail and provides simple visual representations of the elements involved.

3.1 Provenance

The underlying metadata model for all our knowledge assets aligns with the Provenance Ontology proposed by W3C [8]. As seen in Figure 1, and consistent with our “simple is better” approach, we chose to hide some of the complexities of the model and only show relevant Provenance information. It is worth noting that for both Source and Lifecycle transition, we chose to hide the complexities of the underlying models and just display such relations as “simple” pointers. This level of abstraction

will be useful for knowledge authors in that it presents the required Provenance elements while removing details from auxiliary ontologies.

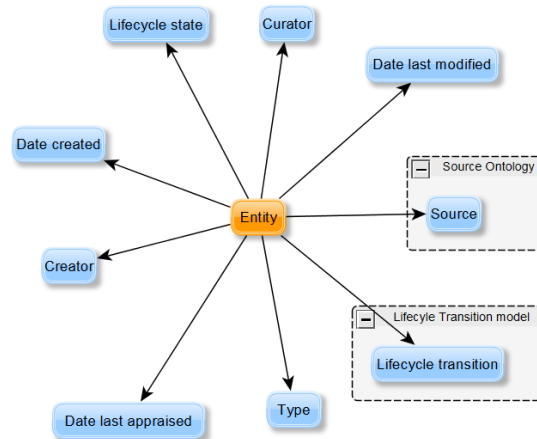


Fig. 1. (Simplified) Provenance metadata in the CKMS metamodel. All entities in the meta-model inherit these provenance properties. Note that the reference to a Source is simplified as a “pointer” to an asset in a Source ontology. Similarly, the Lifecycle Transition is a “pointer” to the current state of the knowledge asset in the Lifecycle Transition Model, again hiding the complexities of the lifecycle model.

3.2 Specific Properties

Figure 2 depicts at a simplified, level, the dependencies among entities in our models for production rule, logic expression, and data element. Even though this depiction might be an over simplification that leaves out most details, e.g. metadata and subtypes – depicted as collapsed nodes, it includes relevant dependencies necessary to understand the components of a CDS production rule. We start with (a) *Production rule*. It consists of an *antecedent* (logic of the rule), and a *consequent*. The *antecedent* is a *logic expression* (b) that could be a *primitive expression* (c) or an *aggregate expression* (d) – a Boolean combination of simple expressions – while consequent represents one or more actions (e.g. notification, assertion, modification or retraction of facts), or some other side-effect. The *logic expression* when evaluated, will return a truth value that will trigger, or not, the execution of the consequent of the rule.

The *logic expression* (b), a type-specific property of the production rule which inherits several properties from its parent type, and declares two additional properties: *Negate expression* and *Topic* (not shown). The former allows for negation of the logic expression, while the latter provides a link to one or more semantic tags from controlled terminologies, such as LOINC, SNOMED, MeSH. As all asset types do, both *Primitive* and *Aggregate expressions* inherit the properties from their parent type, so

negation can occur at all levels in the antecedent regardless of whether it consists of a simple, complex, or nested expression.

Primitive expression (c) type requires a *Data element* (e) with a data type and a comparison operator. Additionally, it contains an optional set operator to build expressions that support checking for element membership in a set (e.g., a problem in the patient’s problem list).

Aggregate expression (d) has a required binary Boolean operator that is restricted to “AND” or “OR” values, as well as a minimum of two *Primitive* or *Aggregate expressions* to support nested expressions.

Data element (e) provides the mechanisms for creating different data element types. All element types include generic and type-specific properties. Generic properties shared among all types include *Topic*, *Reference*, *Purpose* and *Instructions*. *Topic* is of particular importance since it provides the mechanisms for semantic tagging, linking data elements to standard terminologies. This not only allows for a simplified data representation, but also provides the means for validating and standardizing knowledge assets and their interdependencies.

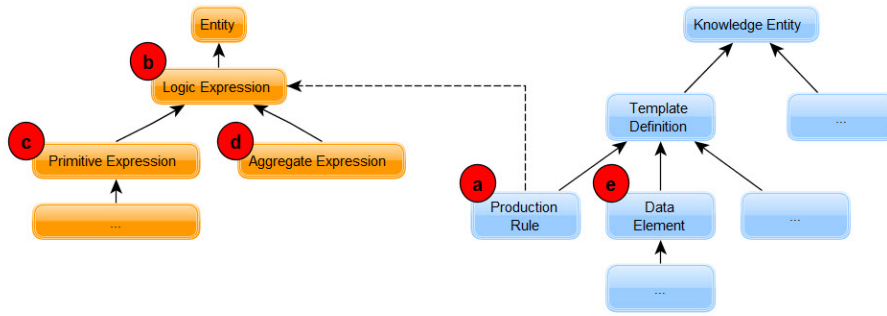


Fig. 2. Ontology models for (a) Production Rule; (b) Logic Expression; (c) Primitive Expression; (d) Aggregate Expression; and (e) Data Elements.

We realize that the description just provided is complex and difficult to grasp. An asset is a subtype of a parent asset (as in the case of a Primitive Expression being a subtype of Logic Expression). An asset has as property an asset from a sister ontology (as in the case of a Production Rule using a Logic Expression). An asset has sister entities or itself as properties (as in the case of Aggregate Expressions). Nevertheless, we believe that again, by removing “spurious” features we can present a simple depiction of complex models that illustrates dependencies among entities (as seen in Figure 2) without overwhelming the user (e.g. knowledge author), and in doing so, we achieve our goal of providing a meaningful, simple and yet powerful representation of a model.

In the next section, we present the “constraints and context dimension” that model the behavior and scope of the production rules.

3.3 Context and Constraints

Capturing context is critical for understanding and handling knowledge. This is particularly true in a clinical setting where knowledge embedded in decision rules often times is tailored to specific scenarios. However, it is also most desirable to preserve the generality of rules, ensuring a high degree of reusability and maintainability.

In addition to Provenance, Constraints are part of the generic properties inherited by all entities in our model. In this section, we will focus our attention on the property *hasConstraint*. The purpose of this property is to delimit rule execution to narrower scopes by restricting it to more specific contexts. As long as we keep this in mind, we can define as many “sub contexts” (more constrained contexts) as needed for a single rule. For example, in its most generic representation, we may have a rule of the form:

If <dataElement><comparisonOperator><thresholdValue> Then <action>

This rule may be applied to a specific laboratory test result, hence replacing *<dataElement>* with a specific value, which will be compared against a *<thresholdValue>* to indicate whether the result is normal or not.

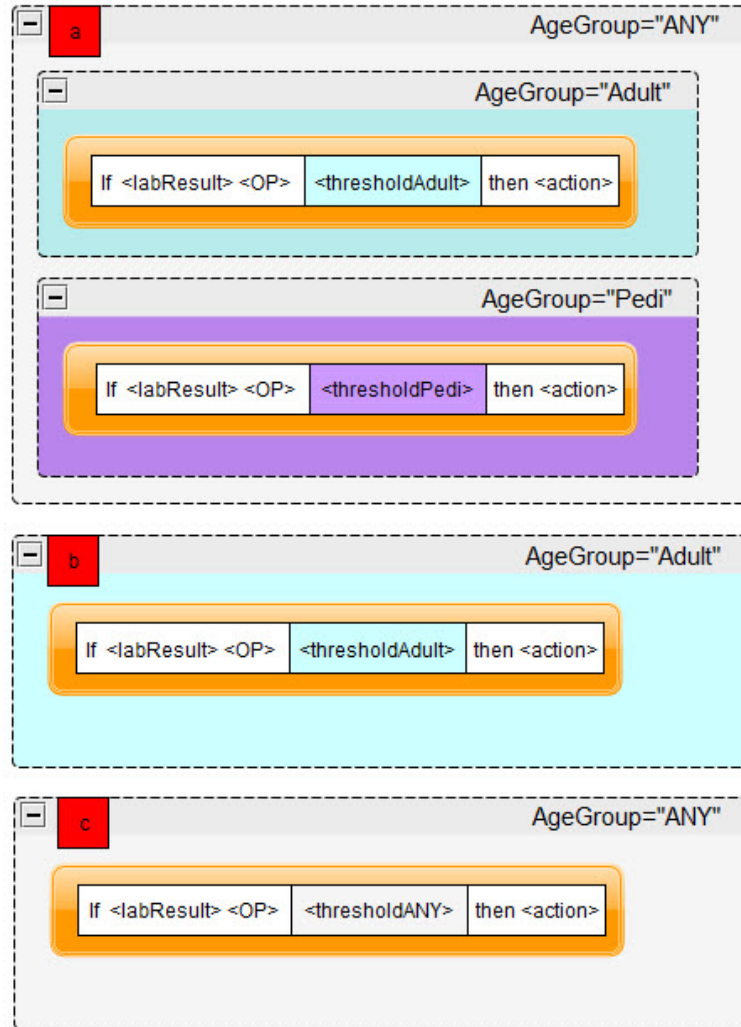


Fig. 3. (a) A production rule with a threshold value constrained to Adult and Pediatric age groups. The logic of the rule remains the same, but each group has its own threshold value; (b) a production rule that is only valid for adults; (c) a production rule that applies to any age group.

This hypothetical rule may apply to both males and females (Gender="ANY") and patients of all ages (AgeGroup="ANY"). The "ANY" value means that such dimension is unrestricted. In some instances, as depicted in Figure 3, threshold value(s) for a production rule may vary depending on the age of the patient. Even though the logic of the rule remains the same, and applies to the overall population, the threshold values are dependent on the age of the patient. This is depicted in Figure 3(a), where two

context-specific threshold values constrain the triggering of the rule to two specific and mutually exclusive age groups: Adult and Pediatric. In Figure 3(b), the rule is constrained (targeted) to the Adult population, and the overall context of the rule itself is restricted to Adult, with no further restrictions on the threshold value. Finally, Figure 3(c) depicts a rule with no age-related constraints; in other words, the rule, and all its components should be able to fire if the condition is satisfied, regardless of the age of the patient population. Further, by specifying such threshold values constrained by age groups in the data definition of the rule expression, we can still model the alerting rule as simply as *if* LabResult < *comparisonOperator* > <thresholdValue> *Then* <action>; where the values assigned to *thresholdValue* are constrained by the context (AgeGroup) where such values apply. Therefore, the logic is the same, but the threshold values are defined by the context. This is consistent with the *context as a box* metaphor. Such metaphor lays the foundation for handling constraints and allows us to manipulate the scope of such rules consistently [11][12].

The narrative in the previous paragraphs raises once again, the issue of describing a complex framework as succinctly as possible. We believe that by coupling the narrative with a simple graphical representation, that removes complexities while presenting relevant features, we conveyed the salient features of this part of the model.

4 Viewing an Instance of a Production Rule

We have created over 150 clinical knowledge assets for abnormal or critical results for chemistry, hematology, and toxicology laboratory tests in outpatient settings using this model. Figure 4 presents an example of a toxicology production rule alerting for abnormal levels of Caffeine. In Figure 4, the production rule (a) has as antecedent “Caffeine greater than 30 mcg/mL” (b), and consequent “Alert level 2” (c).

Figure 4 also shows the antecedent formed by a *Quantity data element* (d). The significance of this is many-fold: the presence of a data element indicates that this information is coming from a clinical system, e.g. patient record, reporting system; the type of the data element (double) indicates information data type (g); the Topic (h) points, in this case, to a LOINC concept (reference terminology) for “Caffeine [Mass/volume] in Serum or Plasma” with code 3422-3 (not shown). By further expanding the nodes, the LOINC concept can provide additional information about the type of concept: quantity; its units of measure: mcg/mL; class it belongs to: toxicology. All this information is readily available just by tagging the *Quantity data element* to the appropriate LOINC code via the Topic semantic tag.

The next level up is the *Primitive expression quantity* (b): “Caffeine > 30 mcg/mL” that holds the *Quantity data element* (d), a comparison operator “>” (e) and a threshold composed of two elements: a value of 30 (not shown) and units of measure mcg/mL (f). Finally, we assign the *Primitive expression* to the antecedent (b), and an *Alert level* to the consequent (c) of the production rule (a): “If Caffeine > 30 mcg/mL then Alert level 2.” This more comprehensive display shows a specific instance of the production rule model. It includes relations to other instances, as well as cardinalities. We have chosen to hide non-relevant information – as depicted by grayed-out nodes

in the graph. The current functionality in our CKMS system allows expanding/contracting nodes to show the required level of information. This functionality enables users to “visually skip” certain details that might be present at any level in the graph, easing understanding of the model.

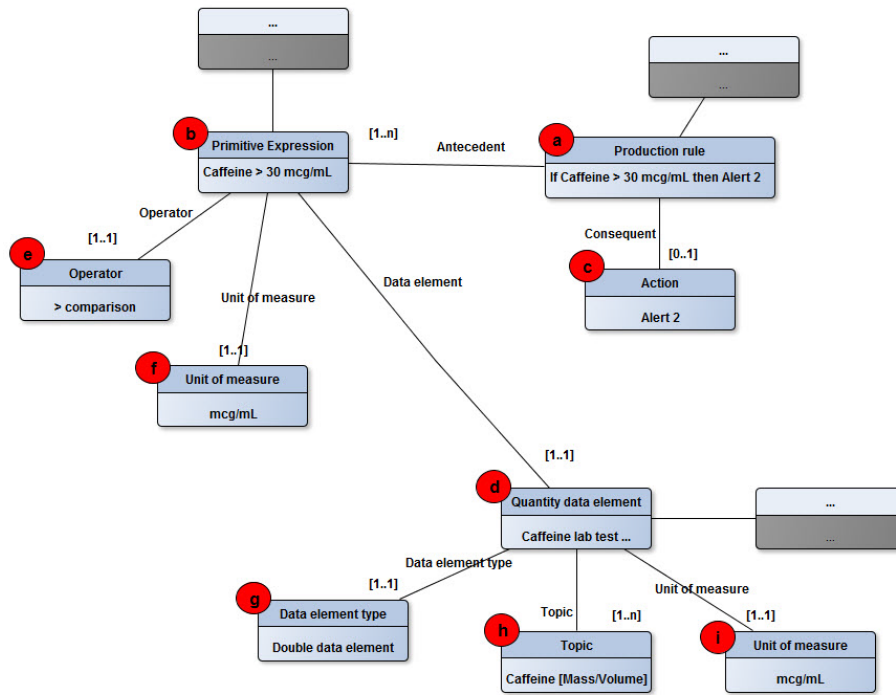


Fig. 4. (a) “If Caffeine > 30 mcg/mL then Alert level 2” instance of the production rule model. Expanded nodes show properties of entity instances, e.g. Production rule, Logic expression, Data element, while also displaying dependencies with reference terminologies and other ontologies as in the case of nodes (f) through (i). Grayed-out (expandable) nodes hide non-relevant information.

5 Discussion

As is the case for all sorts of narratives and depictions, explanations using both words and images have the ability to convey a “story” to highlight the importance of the topic at hand. Coupling narrative with simple visual aids unveils different, simpler aspects of an otherwise complex model.

We posit that even simple, austere depictions may evoke more complex images for the reader, either new or drawn by analogy, which may convey the semantic meaning intended in the first place. Simpler depictions, if well placed, may be more powerful in

communicating the intended explanation, triggering the reader's own mental processes, and interpretations, while enabling a more complete understanding of the presented matter.

We believe the proposed "simple is better" approach applies broadly. Ideas and concepts are more easily assimilated through our cognitive tasks when presented in a simple way. We are more likely to produce our own mental representations and generate new, more complex ideas using such simple models. Elaborate concepts, when presented in their entirety from the beginning, can appear more daunting, deterring us from communicating, exploring and expanding our knowledge. In summary, it is our belief that well-designed data graphics are usually the simplest, and at the same time the most powerful, and can do much more than just substituting words.

Acknowledgments

We would like to thank Andrew B. Phillips, PhD, RN and Aziza D. Daigle, OCM for their valuable assistance in preparing the final version of this document.

References

1. Tufte, Edward R.: *Envisioning Information*. Graphic Press LLC, Cheshire, Connecticut (1990)
2. <http://www.semedy.com/solutions/s-memory-ckms>
3. Rocha RA, Maviglia SM, Sordo M, Rocha BH.: Clinical Knowledge Management Program. In: Greenes, RA, Ed. *Clinical Decision Support. The Road to Broad Adoption*. 2nd Ed. Academic Press (2014) 773-818
4. Garcia, J., Garcia-Penalvo, F., Theron, R.: Modelling relationships among classes as semantic coupling in owl ontologies. In *Proceedings of the 2011 international conference on information and knowledge engineering, IKE 2011* (Vol.1, pp. 22–28)
5. Ware, C.: *Information visualization: Perception for design*. Morgan Kaufmann, San Francisco, CA (2004)
6. F. J. García-Peñalvo, R. Colomo-Palacios, J. García, R. Therón: Towards an ontology modeling tool. A validation in software engineering scenarios. *Expert Systems with Applications* (2012) 39(13):11468–11478
7. Steffen Lohmann, Stefan Negru, Florian Haag, Thomas Ertl: Visualizing Ontologies with VOWL. *Semantic Web* 0 (0) 1
8. PROV-O: The PROV ontology. W3C recommendation. <http://www.w3.org/TR/prov-o/>
9. Sordo M, Maviglia SM, Rocha RA.: Modeling Contextual Knowledge for Clinical Decision Support. Tech Report PHS-2015-MS (2015)
10. Sordo M, Rocha BH, Morales AA, Maviglia SM, Dell'Oglio E, Fairbanks A, Aroy T, Dubois D, Bouyer-Ferullo S, Rocha RA.: Modeling Decision Support Interactions in a Clinical Setting. *Stud Health Technol Inform* (2013) 192:908-12
11. Benerecetti M, Bouquet P, Ghidini C. Contextual Reasoning Distilled. *J Expt Theor Artif Intell*. 12 (2000) 279-305

12. Benerecetti M, Bouquet P, Ghidini C. On the Dimensions of Context Dependence. In: Bouquet P, Serafini L, Thomason RH, Eds. Perspectives on Contexts. CSLI Lecture Notes. Center for the Study of Language and Information/SRI (2007) 1-18

The Landscape Of Philosophy Of Science

Bodil Nistrup Madsen¹, Søren Brier¹, Kathrine Elizabeth Lorena Johansson¹, Birger Hjørland², Hanne Erdman Thomsen¹, Henrik Selsøe Sørensen¹

¹Copenhagen Business School, Copenhagen, Denmark

²Royal School of Library and Information Science, University of Copenhagen, Denmark

Abstract. In Denmark, all higher education programs must include a course on philosophy of science. Therefore, a group of researchers at Copenhagen Business School (CBS) are developing a smartphone application where information about central theoretical paradigms and concepts from philosophy of science can be visualized and disseminated in an easily accessible and systematic manner. This will be achieved by entering structured knowledge about concepts from philosophy of science in both Danish and English into a terminology and knowledge base which will provide the opportunity to “navigate in conceptual landscapes” (here used metaphorically for terminological ontologies) in the same way as we navigate in maps. The result of the project will be a tool that can help students in their studies and support their information retrieval. The project is based on existing technologies and research in knowledge organization and knowledge management. In this paper we will present the first version of a terminological ontology of central paradigms.

Keywords. Terminological ontologies, Philosophy of science, Ontology app.

1 Introduction

Philosophy of science has been a mandatory undergraduate course in all university programs in Denmark for at least 10 years. Characterized by a high level of abstraction, the discipline presents major challenges to students, who are used to thinking in concrete terms. The attempt to remedy this by introducing a general university preparatory course in high school has had little effect, as it is not consistently taught in all secondary and higher preparatory schools. From many years of experience with planning and teaching courses in philosophy of science in multiple disciplinary programs, it has become clear that the central challenge for students is to obtain an overview of the many epistemological and paradigmatic concepts, as well as a general understanding of how these concepts challenge our common sense understanding of reality. The students are constantly demanding accessible overviews and explanations, and need access to summaries and short, consistent descriptions of the different paradigms in a way that can support their learning process. There are various attempts to provide students with this in the avalanche of textbooks in philosophy of science published over the last ten years in Denmark alone, but none of them present an adequate

solution. When they are clear and thorough, they are often very long and cumbersome to read; if they are designed to be brief, they are either not consistent or prone to errors of simplification.

In response to the extensive discussions about how to use the new social media technologies and hypertext forms that our digital native students use extensively, the project carried out in the research group for Representation, Organization and Communication of Knowledge (ROCK) at CBS aims to strengthen the teaching situation and to support the students' motivation by creating a platform that fits directly into their current life. This is why the ROCK group will develop an ontology app for a smartphone, thus providing students with easy access to concepts organized in terminological ontologies and provided with clear and consistent definitions, which ensure that academic concepts are understood and applied correctly. Thus, the aim of the app is two-fold: to support the students' motivation for research based learning, and to level with international standards in order to constitute a common basic ground within the field of philosophy of science.

2 Terminological ontologies

The ontology app will consist of an advanced term base based on the theory of *terminological ontologies*, c.f. Madsen, Thomsen and Vikner (2004). The principles of terminological ontologies are based on the directions concerning *concept systems* in ISO standards, such as ISO 704:2009, and are developed by a research group at the Dept. of International Business Communication, CBS, in the CAOS project (1998-2007), which aimed at semi-automatic development and validation of terminological ontologies. In Madsen, Thomsen and Vikner (2004) we describe a number of constraints and principles which apply to terminological ontologies. On the basis of these principles, the DANTERMcentret has developed a terminology and knowledge management tool, i-Term[®], which comprises an ontology modeling module, i-Model. We use this tool to store and visualize the results of the concept clarification carried out in the current project. The concept modeling in i-Model is based on user input, and has no automatic consistency checking facilities.

As an introduction to the description of the current project we present some central concepts related to terminological ontologies. The first example originates from the terminology work carried out by a number of working groups established by the Danish Council for Health Terminology. Figure 1 presents a small extract of the terminological ontology for disease prevention, re-created using the concept modeling module i-Model.

Basically, terminological ontologies are terminological concept systems enriched with characteristic features in the form of attribute-value pairs based on Carpenter's "Typed Feature Theory" (Carpenter 1992). Concept systems contain concepts and relations between these. A concept is the meaning of a term (or intension), and it is reflected by a class of objects (the extension of the term). A concept system in terminology does not contain classes or objects/instances, but the class of objects corresponding to a given concept may contain 0,1 or more objects. Concept systems and

terminological ontologies are concerned with intensional semantics, rather than the extensional semantics underlying e.g. First Order Logic, c.f. Madsen and Thomsen (2009: 542).

The concepts in a terminological ontology are described by means of semantic information, i.e. their mutual relations and the characteristics, given in a formalized form, though still understandable by humans. This information is used for both building and validating terminological ontologies, and based on these, intensional definitions may be developed. For each concept one or more synonymous terms with associated information (references, examples, phraseology, notes etc.) may be registered and be accessed by double clicking. In Figure 1, each yellow box corresponds to a concept represented by the preferred term for that concept. Lines between concepts correspond to type relations (in terminology known as generic relations). Other relations (part-whole relations, associative and temporal relations) may be represented with different line types. The characteristics of the concepts are presented below the concepts as feature specifications in the form of attribute-value pairs, e.g. *TARGET GROUP: population*. Coordinate concepts (concepts with the same superordinate concept) contain characteristics with the same attribute, but different values. Such attributes are referred to as *dimensions*. The dimensions are, at the same time, *subdivision criteria* according to terminology theory (Arntz & Picht 1989:83). The subdivision criteria in Figure 1 (white boxes with text in capital letters) illustrate that the three coordinate concepts 1.1 to 1.3 differ with respect to *target group*, while the three concepts 1.4 to 1.6 differ with respect to *phase in clinical course*. The subdivision criteria help the user to understand the meaning of the concepts, give a good overview and help the terminologist in writing consistent definitions.

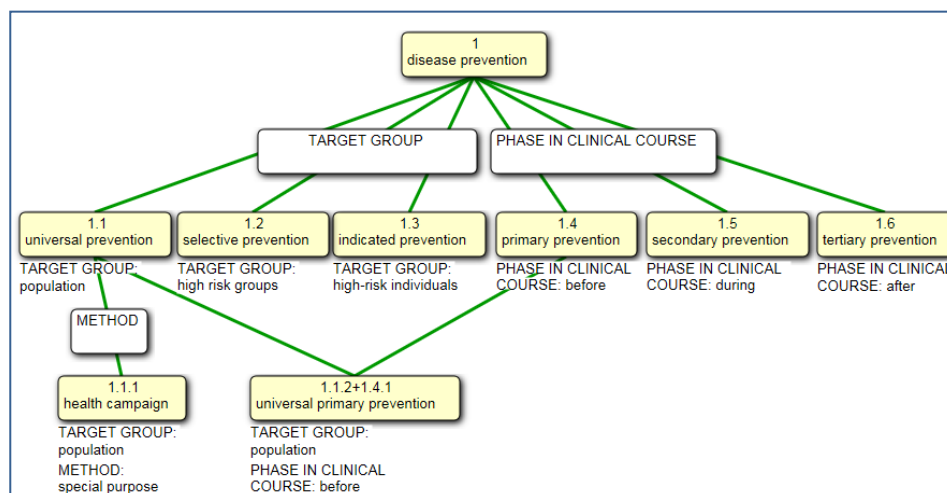


Fig. 1. Extract of the terminological ontology for disease prevention

In the following a brief description of two principles of terminological ontologies, which are relevant for the ROCK ontology, will be given.

The first principle states that subdivision criteria should be chosen in such a way that 1) all subordinates are covered and 2) no concept belongs under more than one criterion of subdivision, c.f. Madsen & Thomsen (2015). This implies that in some cases, the subdivision criterion must be chosen among several possibilities. In Figure 2, the concept *selective prevention* is characterized by being targeted at high-risk groups and carried out by health care professionals in risk environments, whereas *universal prevention* is characterized by being targeted at the entire population and carried out by public authorities (typically) in schools. This results in three dimensions, TARGET GROUP, ARENA and AGENT, which are all potential subdivision criteria, but one of them has to be chosen, in order to comply with the principle. In this case, it can be argued that the ARENA and the AGENT follow from the choice of TARGET GROUP, i.e. who can get into contact with the target group and where, and therefore TARGET GROUP must be chosen as the subdivision criterion.

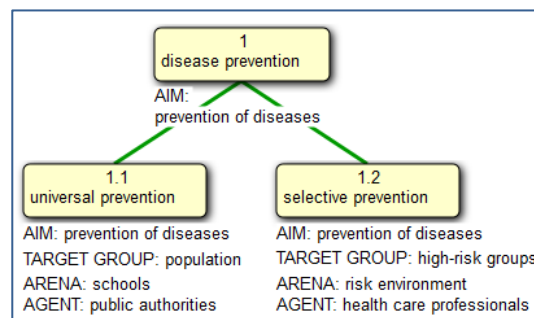


Fig. 2. Several potential subdivision criteria

The second principle states that an attribute may only be associated with one value in a feature structure (a combination of two or more feature specifications on a concept is called a feature structure). In Figure 1, it would for example be a violation of this principle to insert a concept *primary secondary prevention* with two superordinate concepts under the same subdivision criterion, *primary prevention* and *secondary prevention*. In this case the attribute PHASE IN CLINICAL COURSE would be associated with two values in the feature structure: *PHASE IN CLINICAL COURSE: before* and *PHASE IN CLINICAL COURSE: during*. This is *illegal polyhierarchy*.

Figure 1 contains an example of a *legal polyhierarchy*: the concept *universal primary prevention* with two superordinate concepts under two different dimensions (TARGET GROUP and PHASE IN CLINICAL COURSE). In the case of polyhierarchy, the combination of the feature specifications inherited from the superordinate concepts distinguish the concept from other concepts, and the definition should comprise both characteristics. It should be noted that it is possible to introduce the concepts *selective primary prevention* and *indicated primary prevention*, but not a concept *universal tertiary prevention*. Tertiary prevention aims to soften the impact of an ongoing illness or injury that has lasting effects, and the target group of tertiary prevention is individual patients, not the whole population. Therefore *universal tertiary prevention* is a concept which does not exist in reality.

In Figure 3 we present an example from the domain of enzyme chemistry. The four subordinate concepts to the concept *reversible inhibition* differ with respect to two feature specifications with the attributes MICHAELIS CONSTANT and MAXIMUM RATE. Here it is not possible to choose one dimension as a subdivision criterion, which appears clearly from the values of the attributes, c.f. Damhus et al. (2009).

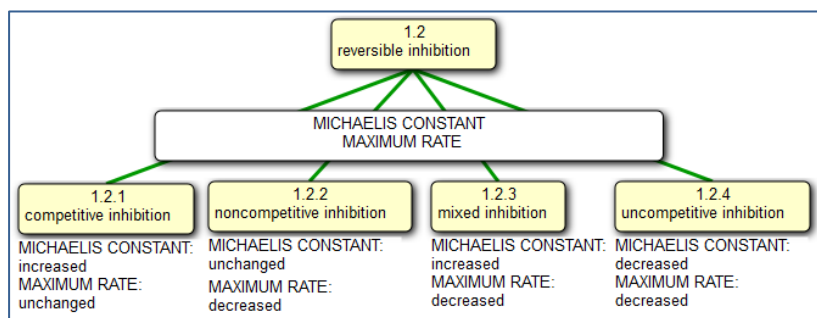


Fig. 3. Concepts delimited by a combination of characteristics

In the ontology in Figure 4, a layer of extra concepts was therefore introduced: three concepts (1.2.1 to 1.2.3) that differ with respect to *MICHAELIS CONSTANT* and two concepts (1.2.4 to 1.2.5) that differ with respect to *MAXIMUM RATE*. These concepts are non-lexicalized and are maybe not important for the purpose of concept clarification. However, if one wants to adhere to the principles of terminological ontologies for formalizing the ontology with a view to consistency checking, this layer of concepts is necessary, c.f. Madsen and Zambach (2009).

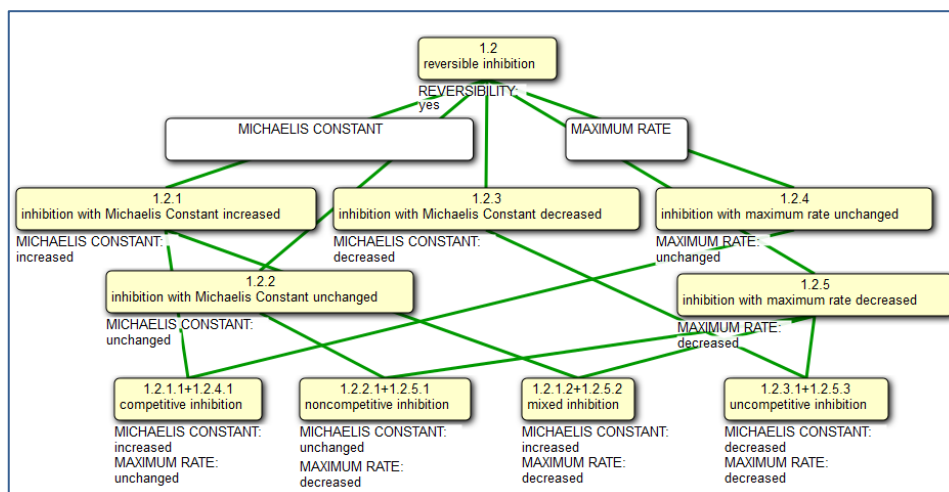


Fig. 4. Diagram with non-lexicalized concepts

3 The ROCK ontology

3.1 Overall structure

In the appendices, we present various views of the terminological ontology of central scientific paradigms, the ROCK ontology. First we describe the overall structure of the ontology, and then we discuss the characteristics used to describe the paradigms. Appendix 1 presents an overview of the ROCK ontology with temporal and selected associative relations. The full set of relations does not fit into the space allowed here. Appendix 2 presents another overview with only type relations.

The big problem has been to find a transdisciplinary framework for ordering the paradigms. We needed to include the classical paradigms that we teach all students and also some of the new and often more complex ones. As a more general history of ideas framework we chose *modernism* and *postmodernism*. We then chose to frame the paradigms on a scale between *classical* and *complex paradigms* (such as *positivism* and *Cyber semiotics*) in that it is typical for the early paradigms to attempt a reduction to one system, where the later ones are usual more complex and attempting to encompass quantitative and qualitative aspects of science (broadly understood as *Wissenschaft*). Then we chose *realism* versus *non-realism* as central concepts to distinguish all different types of *constructivism* and *relativisms* from those who had concepts of truth as essential. We realized that *dialectical realism* represented a special process form of *realism* and inserted this in the middle between the two extremes.

3.2 The nature of the characteristics

Figure 5 presents an extract of the ROCK ontology with characteristics. Here the complex nature of the domain of philosophy of science becomes very clear.

Ideally the values in the feature specifications should be concepts themselves, i.e. they should be short like e.g. the value of the attribute ONTOLOGY on the concept *Actor-network theory* ('constructivistic realism'), and not correspond to sentences, such as the value of the attribute EPISTEMOLOGY on the concept *systems theory* ('objective knowledge can be found by understanding systemic structures, and ...').

Furthermore, all paradigms are described by means of a combination of characteristics, which violates the above mentioned constraint that no concept should belong under more than one criterion of subdivision. This means that the characteristics come in clusters, as illustrated in Figure 2. One solution to this is to choose one dimension as subdivision criterion and consider the other characteristics as dependent on the one in question. This attribute could maybe be PURPOSE, since all paradigms have different values to this attribute. Another solution could be the attribute ONTOLOGY, which generally has shorter values, but in two cases the same value is used on two concepts, e.g. the value 'relativism', which is found on the concepts: *postmodernism* and *constructivism*.

A third solution could be to consider all paradigms as concepts in a poly-hierarchical structure, where a number of 'non-lexicalized' concepts could be introduced in order to make the underlying polyhierarchical structure explicit, such as e.g.

paradigm with ontology based on constructivistic realism (c.f. Actor-network theory), paradigm with ontology based on 3rd person structural realism (c.f. systems theory), etc. and paradigm with the purpose of examining varied translations in heterogeneous networks, paradigm with the purpose of generating a hierarchical and general theory of systematization, etc. The paradigm Actor-network theory would then be a subordinate concept to two concepts: paradigm with ontology based on constructivistic realism and paradigm with the purpose of examining varied translations in heterogeneous networks. Having accepted that this polyhierarchical structure exists, it is possible to also accept that each paradigm is described by means of a combination of characteristics. It would definitely not be helpful for students to make this structure explicit, unless it would be possible to shorten the values of the attributes and thereby also to introduce much shorter terms for the non-lexicalized concepts.

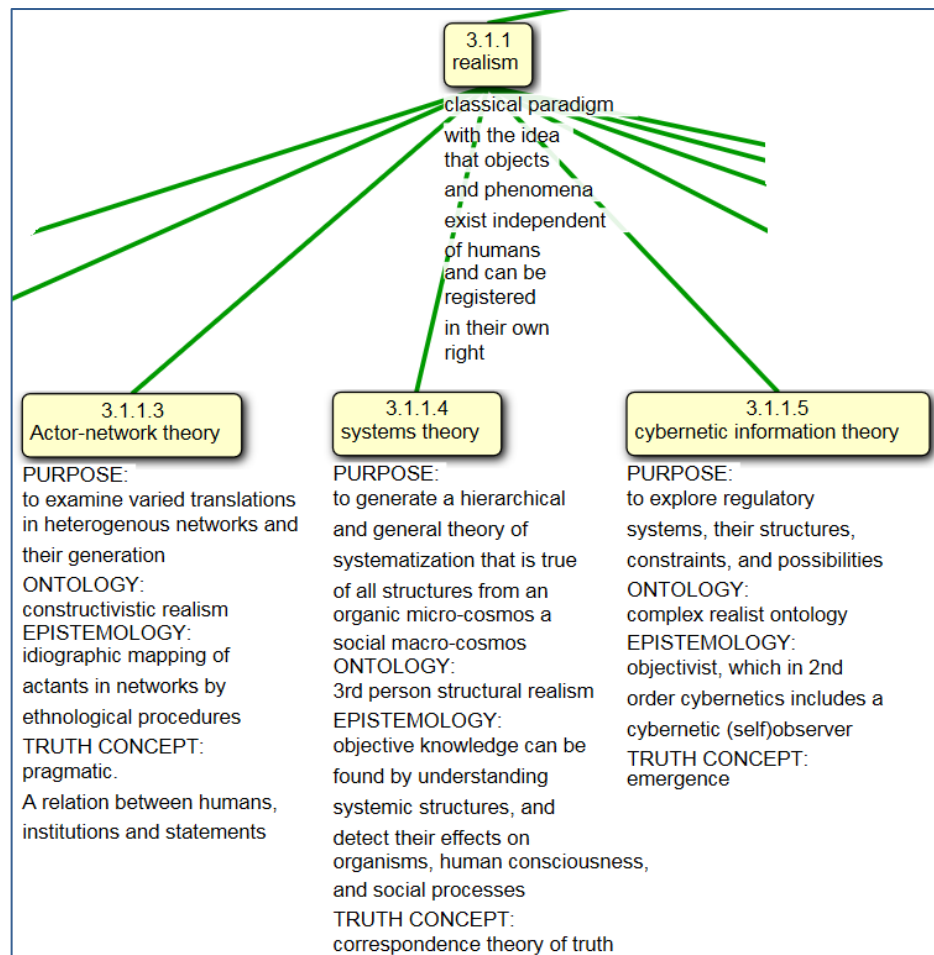


Fig. 5. Diagram with non-lexicalized concepts

4 The view of philosophy of science behind the ROCK model

There are many different perspectives on philosophy of science, which makes it difficult to build a single ontology, rather than several ‘competing’ ontologies, based on the different views, c.f. section 5 Perspectives. At this point we have developed one, aimed primarily at students in specific CBS courses of philosophy of science. These courses use the text book Brier (2006), and hence the view represented in this book is reflected in the ROCK ontology.

The logical positivists were right in that scientific theories should be as much as possible able to predict observable reality. But scientific theories, although they do get corroborated, as Popper (1972) calls it, cannot be verified or proven from correspondence with empirical observations alone. Empirical testing does strengthen our belief in them though. But simple correspondence between word and object (or sentence and state of affairs) provides very little explanatory value. Thus, there are unavoidable underlying ontological, epistemological and axiological commitments in holding a term or sentence to be true. This brings us part of the way to Kuhn’s ideas about paradigms, or what other researchers call ‘research programs’, ‘schools’, ‘isms’ or ‘epistemological positions’. In this paper we consider these terms near-synonyms, since a further distinction will require more research. Here we will just remark that they all need to have an empirical methodological aspect, like for instance phenomenology. We do not deal with purely theoretical philosophical programs, like for instance NeoKantianism.

Observation is always made on the basis of a problem interest, Popper points out. Thus, observations are never really disinterested and objective in themselves. These background interests and assumptions of the researcher should, therefore, also be reflected on and stated clearly for others to evaluate the knowledge generated, as also Gadamer (1975) underlines. All theories have presumptions about the nature of reality, cognition and knowledge, from which their methods, scientific objects, and subject areas are defined. This is what Kuhn (1970), in the second version of his paradigm theory, describes as the *disciplinary matrix*.

Although the presently accepted scientific theories cannot be shown to prove the truth about the world or reality to us, they nevertheless contain a lot of tested new knowledge about parts or aspects of reality. Our knowledge is without doubt growing like an island that is expanding in all directions in a sea of potential knowledge as Kuhn (1970) points out.

5 Perspectives

The modeling of an ontology for theory of science has been going on for more than a year and was expected to be a rather simply enterprise, since it was meant to cover only key concepts and paradigms relevant to bachelor students having a one term course. This course, which is taught by a good many teachers, leads to an exam, where all hand-ins should be graded according to the same criteria irrespective of who taught the course. On this background it could be expected that an agreement on the

key concepts and the presentation of these in the form of an ontology could be reached.

Nevertheless, this was not the case. This fact lead to an idea for a future development which would make use of the flexibility of the i-Term system. The idea consists in allowing the same paradigms to appear in competing ontology structures built according to different perceptions of the teachers, which it would be perfectly possible to implement in i-Term. A given student would then be free to choose the relevant view represented by his/her teacher when consulting the knowledge base.

This way of combining competing ontologies would also solve problems occurring when unavoidable changes of teachers appear. Furthermore, a pilot implementation could pave the way for inventing a model which would solve similar problems occurring very frequently in other contexts, for example when multinational teams must collaborate but experience problems rooted in different cultural and educational perceptions in situations, where no one has the power to force a particular view on to all team members.

It should be emphasized that the research carried out by the ROCK group is ongoing, and that the ROCK ontology should be compared to the results of similar research, c.f. for example Iivari, Hirschheim & Klein (2001).

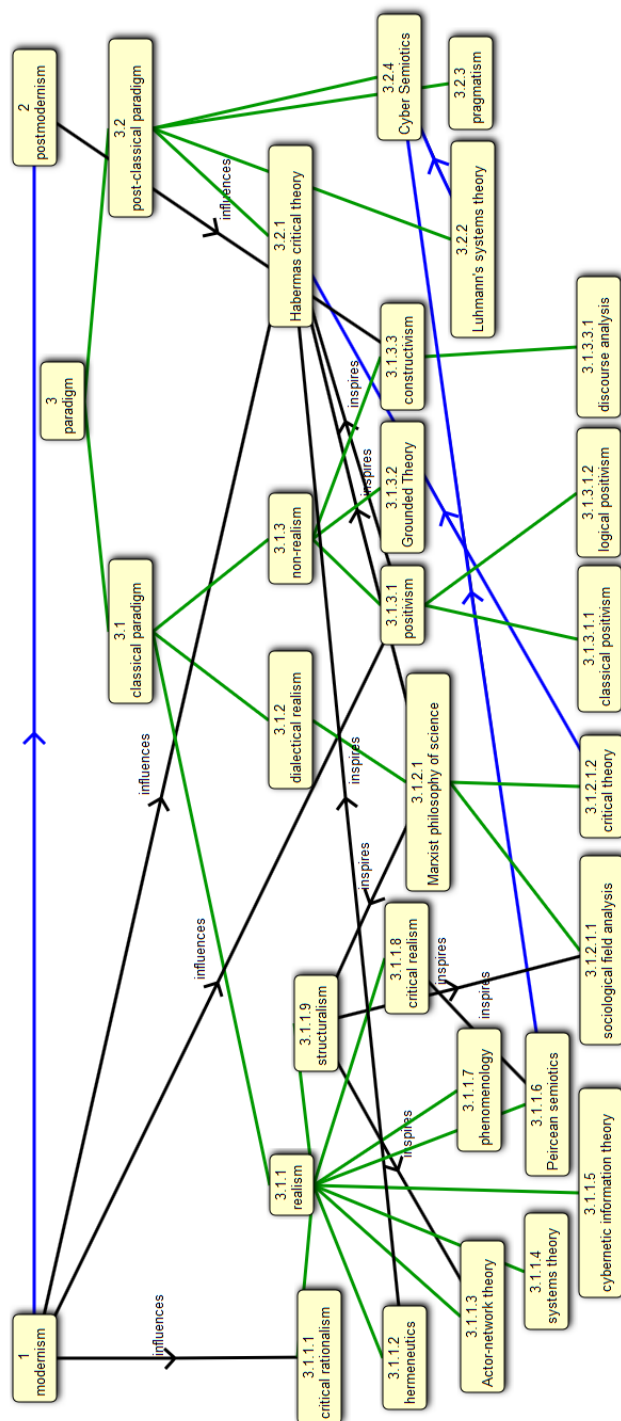
6 Conclusion

The terminology and knowledge base resulting from this project is relevant for students in all higher education in Denmark and at an international level. It will be accessible by means of an app, which can be reused for other subjects than philosophy of science. Philosophy of science is a very complicated domain, and, as it has become clear in our paper, it is difficult to give brief descriptions of the concepts. However, the terminological ontology with the characteristics in the form of feature specifications gives a much better overview of the paradigms than long texts. The app, which will be used for visualizing terminological ontologies in a smart way, is still at a prototype level, but when finished it will make it possible for students to navigate in the landscape of philosophy of science using their smartphone.

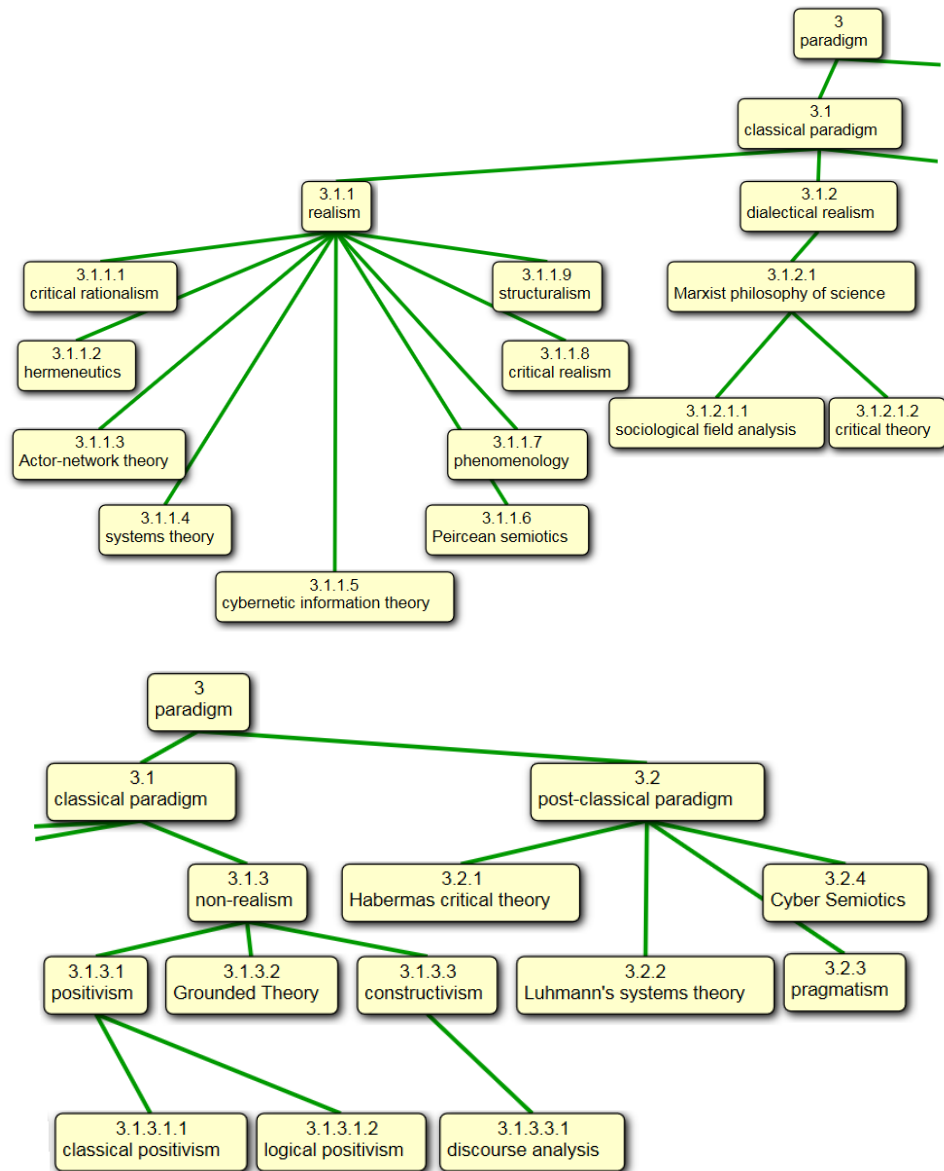
Acknowledgements

We wish to thank our colleagues who participated in the work on the ROCK ontology and development of the prototype of the ontology app: Per Durst-Andersen, Daniel Barratt, Zhou Liqian, Barbara Dragsted, Louise Pram Nielsen, Leen Boel and Radu Dudici.

Appendix 1: Extract of the ROCK ontology with temporal and selected associative relations



Appendix 2: Extract of the ROCK ontology (only type relations)



References

- Arntz, Reiner & Heribert Picht. (1989). *Einführung in die Terminologearbeit*. Hildesheim, Georg Olms AG.
- Brier, Søren. 2006. *Informationsvidenskabsteori*, København: Forlaget Samfundslitteratur.
- Carpenter, Bob. 1992. *The Logic of Typed Feature Structures*. Cambridge: Cambridge University Press.
- Damhus, Ture, Peder Olesen Larsen, Bodil Nistrup Madsen & Sine Zambach. 2009. Consistency and Clarity in Chemical Concepts. How to Achieve a Codified Chemical Terminology – A Pilot Study. *Chemistry International, Volume 31 No. 5*, 6-11.
- Gadamer, Hans-Georg. 1975. *Truth and Method*, New York: Seabury Press.
- Iivari, Juhani, Rudy Hirschheim, & Heinz K. Klein. 2001. A dynamic framework for classifying information systems development methodologies and approaches. In: *Journal of Management Information Systems*, 17(3), 179-218.
- ISO 704:2009. *Terminology work - Principles and methods*. Geneva: International Standards Organisation.
- Kuhn, Thomas. 1970. *The Structure of Scientific Revolutions*, 2nd enlarged ed. Chicago: The University of Chicago Press.
- Madsen, Bodil Nistrup & Hanne Erdman Thomsen. 2009. Terminological concept modeling and conceptual data modeling. In: *International Journal of Metadata, Semantics and Ontologies (IJMSO) Vol. 4 No. 4*, 239-249. Online: <http://www.inderscience.com/link.php?id=29228>.
- Madsen, Bodil Nistrup & Hanne Erdman Thomsen. 2015. Concept Modeling vs. Data modeling in Practice. In: *Handbook of Terminology*. Eds. Hendrick J. Kockaert; Frieda Steurs. Vol. 1 Amsterdam: John Benjamins Publishing Company, 250-275.
- Madsen, Bodil. Nistrup, Hanne Erdman Thomsen & Carl Vikner. 2004. Principles of a system for terminological concept modeling. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 1, 15-18.
- Madsen, Bodil Nistrup & Sine Zambach. 2009. Applying terminological methods and Description Logic for creating and implementing an ontology on inhibition. In: *Proceedings of the International Conference on Knowledge Engineering and Ontology Development. KEOD09*. Madeira, INSTICC.
- Neurath, Otto. 1983. *Philosophical Papers 1913–1946*, R.S. Cohen and M. Neurath (eds.), Dordrecht: Reidel.
- Popper, Karl. 1972. *Objective Knowledge: An Evolutionary Approach*, Oxford: The Clarendon Press.

Target Users' Diagrammatic Reasoning of Domain-Specific Terminology

Louise Pram Nielsen

Dept. of International Business Communication, Copenhagen Business School

lpn.ibc@cbs.dk

Abstract. In this paper, we investigate target users' diagrammatic reasoning in a controlled experiment, where participants were asked to search for information in a dual visualization comprising of a concept-oriented graphical (diagram) entry and a corresponding textual (article) entry. During the experiment, users' visual attention was recorded by means of eye-tracking technology. We chose professionals as participants and taxation as our exploratory domain. We show that diagrammatic reasoning is effective and improving on questions related to diagrams only (so-called D-questions). However, significantly longer response time was needed to produce correct answers to D-questions compared to the questions related to articles only (so-called A-questions) as well as questions related to both diagrams and articles (so-called DA-questions). Hence, diagrammatic reasoning of the D-questions is the least efficient compared to A- and DA-questions.

Keywords: domain-specific terminology · dual visualization · diagrammatic reasoning · eye tracking

1 Introduction

Terminological ontologies allow for the visualization of concepts, relations and characteristics in the graphical format, which renders possible target users' acquisition of domain-specific terminology and knowledge. It has been widely recognized that effective ways of visualizing ontologies vary [1]. Indeed, a great potential for extending the conventional textual format of e.g. dictionaries with graphs exists (see e.g. [2]). Electronic dictionaries provide for a multimodal representation of meaning [3], which in the case of complementary multiple representations is expected to facilitate the learning of users by reducing the overload of single representations [4]. Unsurprisingly, enhancing target users' access to data constitutes the primary concern of future lexicography [5].

Terms express an underlying concept [6] and it should be stressed that terms and their underlying concepts are members of different semiotic systems [7]. Terminological ontologies are the result of an analysis of characteristics, which the terminologist obtains from the specialized discourse either by means of extraction of term candidates, relations and definitions occurring in specialized texts or by consulting the subject

experts of a particular domain [8]. The linguistic units (or so-called textual cues) extracted from the specialized discourse are modelled into domain-specific concepts by formal feature specifications of attribute-value pairs [9]. In other words, terminological ontologies are the result of introducing formal ontology [10] to the practice of terminology work [11].

The textual information collected by the terminologist may be visualized in a conventional concept article and assigned to the relevant part of the terminological ontology visualizing the underlying concept including the relevant super-, side- and subordinate concepts (cf. figure 1). This duality between textual and graphical visualization of concepts and terms is key to the experiment in our research as we provide target users with access to a dual-entry mode, which we expect will facilitate target users' acquisition of knowledge.

The purpose of this paper is to inquire into the diagrammatic reasoning of potential target users of the proposed dual-entry mode using professionals as participants and taxation as pilot domain. In particular, the research question is whether domain-specific terminology and knowledge can be conveyed to target users by means of diagrams. The paper is outlined as follows: In section 2, we describe the methods applied and data collected in the experiment. In section 3, we describe the regression technique and present the results. In section 4, we conclude and indicate directions for future work.

2 Method and Material

The research question is answered by applying experimental eye-tracking methods, which are well-suited to isolate explanatory effects underlying participants' behaviour. In the field of cognitive psychology, established research methods are dominated by the experimental design and procedures because the controlled tasks allow for objective performance measures from which users' underlying cognitive processing may be inferred [12].

The experimental procedure was as follows: Prior to the dual-entry-mode experiment, participants answered a background questionnaire and completed a number of tasks designed to reflect their level of expertise. Then the actual eye-tracking experiment was conducted, and finally, a structured retrospective interview (15 questions) produced auxiliary subjective data on users' perceptions and preferences in addition to the objective eye-tracking data. Below, we motivate the chosen sampling of participants (cf. section 2.1), dual-entry modes and multiple-choice question format (cf. section 2.2), and eye-tracking methods (cf. section 2.3).

2.1 Participants

We limited our research to professional potential target users of a terminology and knowledge bank providing users with the visualizing of concepts in the proposed dual-entry mode (cf. section 2.2). Professionals were members of e.g. legal, financial or administrative staff with advanced working tasks, who would consult the terminology and knowledge bank in connection with their work. Using professionals as target users

are most likely ensuring that our sample represent the full scale of expertise ranging from low to high. This would not necessarily be the case for non-professionals.

We aimed for an unbiased sample of participants representing the potential target users across age and gender, as well as the proposed set of background expertise variables, including participation, motivation, exposure to relevant specialized discourse, and education. In particular, we used participation interpreted as work place as the primary selection criteria. Therefore, 20 participants were staff members of the Danish Customs and Tax Administration (SKAT), while 20 participants were working outside SKAT. In total, 40 volunteers, 23 females (mean age 41.7) and 17 males (mean age 44.0) were sampled from the relevant population of professional potential target users of a domain-specific terminology and knowledge bank. Prior to the 40 experiments, five volunteers (one from SKAT and four from CBS) ran pilot versions causing minor adjustments to the experimental design.

2.2 Dual-Entry Modes and Multiple-Choice Questions

We designed a dual-entry-mode template displaying target concepts in text and graph inside the stimulus space of the screen just below the multiple-choice questions (cf. figure 1). The textual entry mode is displayed as a bilingual written article in tabular format with two columns and ten rows, while the graphical entry displayed information in a conceptual diagram including 5-7 concepts (nodes) structured in three levels of subdivision criteria. This choice of template design allowed for very little difference in the layout, which ensured comparability across the eight different dual-entry modes.

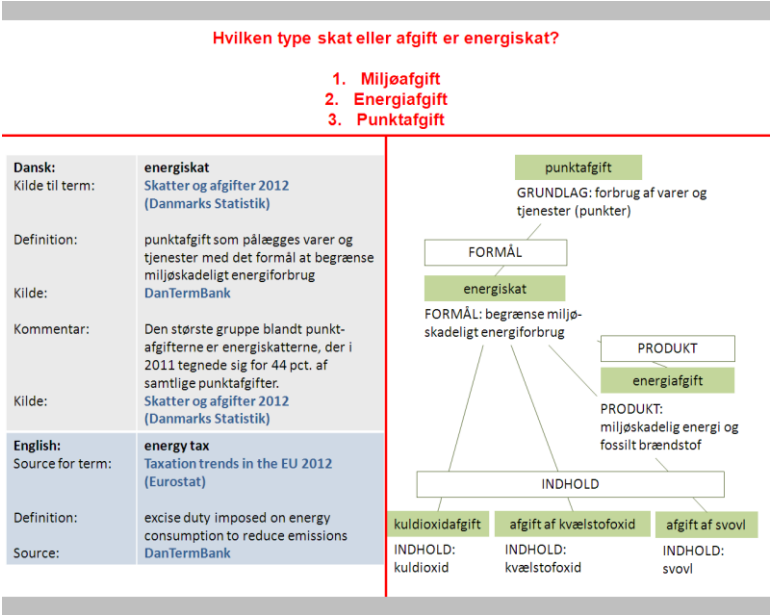


Fig. 1. Example of dual visualization in the case of the target concept “energy tax” (*energiskat*), question type DA and the diagram displayed to the right. Three areas-of-interest

(AOIs) are introduced, one for the question area and one for each of the two types of entry mode. Translation into English: **Question area:** (*Hvilken type skat eller afgift er energiskat?* 1. *Miljøafgift*; 2. *Energiafgift*; 3. *Punktafgift*.) What type of tax or duty is energy tax? 1: Environmental duty; 2: Energy duty; 3: Excise duty. **Diagram entry:** Superordinate (*punktafgift*): 'excise duty' with subdivision criteria: TAX BASE: consumption of goods and services. Entry (*energiskat*): 'energy tax' with subdivision criteria: PURPOSE: limiting environmentally damaging energy consumption. Subordinate (*energialgift*): 'energy duty' with subdivision criteria: PRODUCT: environmentally damaging energy or fossile fuels. Subordinate (*kuldioxidafgift*): 'carbon dioxide tax' with subdivision criteria: CONTENT: carbon dioxide. Subordinate (*afgift af kvælstofoxid*): 'duty on nitrogen oxides' with subdivision criteria: CONTENT: nitrogen oxides. Subordinate (*afgift af svovl*): 'duty on sulphur' with subdivision criteria: CONTENT: sulphur. **Article entry:** Row 1: Danish: energy tax. Row 2: Source for term: Taxes and Duties 2012 (Statistics Denmark). Row 3: Definition: excise duty on goods and services with the purpose of limiting the environmentally damaging energy consumption. Row 4: Source: DanTermBank. Row 5: Comment: The largest subgroup of the excise duties is energy taxes, which constituted 44 per cent of total excise duties in 2011. Row 6: Source: Taxes and Duties 2012 (Statistics Denmark).

We applied the multiple-choice-question format in order to keep the participants' answering process as well as the evaluation process as simple as possible avoiding the time-consuming interpretation and coding of non-restricted answers. In particular, participants were asked to pick one of the numbers "1", "2" or "3" to represent the correct answer. Moreover, the multiple-choice-question format constituted the clear advantage that we were not only providing participants with correct answers, we also provided (plausible) wrong answers. This means that the format forced participants to carefully consider plausible alternatives, which may closely resemble a realistic concept-clarification or knowledge-acquisition user situation.

In total, the experiment contained 48 questions (i.e. the corresponding variable 'trial number' ranges from 1 to 48). The experimental design included eight target concepts: Four belonging to indirect taxation: "energy tax" (*energiskat*), "motor vehicles tax" (*afgift af motorkøretøj*), "green tax" (*grøn afgift*) and "excise duty" (*punktafgift*); and four belonging to direct taxation: "middle-bracket tax" (*mellemskat*), "land tax" (*ejendomsskat*), "personal income tax" (*personskat*) and "direct tax" (*direkte skat*).

Each target concept was visualized as dual-entry-mode pair and assigned to a block of six questions (cf. Appendix D in [13] for the dual-entry modes of each target concept) with three question types, where the answer was found in the article only (denoted A-question), the diagram only (denoted D-question) or either diagram or article (denoted DA-question). In other words, each question type had two conditions allowing the diagram to be displayed to both the left and right, and with three available answers for each question.

The six questions belong to the following categories (cf. table A1): The first diagram-based question (denoted "D1") concerned sub-ordinates and the second diagram-based question (denoted "D2") concerned sub-division criteria. The first article-based question (denoted "A1") concerned equivalence and the second article-based question (denoted "A2") concerned comments. The first diagram-and-article-based question (denoted "DA1") concerned super-ordinate and the second diagram-and-article-based question (denoted "DA2") concerned attributes.

In other words, we manipulated the questions carefully to ensure that participants consult both concept diagrams as well as articles. Research on multimedia has shown that participants have a preference for their most familiar representation [4] and to avoid any biases induced by participants' preferences, the experiment was randomized at three levels. Thus, display side of the diagram, question type and target concepts were presented in randomized order.

2.3 Eye-Tracking Methods

Eye-tracking methods are widely used in dictionary research (see e.g. [14]) and we chose to apply eye-tracking technology in the experiments to inquire into users' visual attention [15] as evidence for the underlying cognitive processing of the proposed dual-entry-mode stimuli (cf. figure 1).

We apply the famous eye-mind hypothesis [16] implying that there is presumably no appreciable lag between fixation and cognitive processing. However, we should expect a minor discrepancy between participants on-screen eye movements and their cognitive processing, as we cannot be entirely sure that what participants look at is also triggering their processing [17]. In our view, it is possible to overcome this discrepancy if we apply mixed methods and collect auxiliary data, which allows us to support and enrich the interpretation of the results indicated by the eye-tracking data. In other words, we combine the distinct quantitative and qualitative methods to gain access to the cognitive processes of participants [18].

In practice, we used a remote SensoMotoric Instrument (SMI) eye tracker, which supports gaze sampling rates of 50 Hz to record participants' on-screen eye-movements during the experiment. We applied fixation thresholds above 200 ms and our primary areas-of-interest (AOI) corresponded to the visualization modes (text or graph) constituting a large part of the screen (cf. the red frames of figure 1).

3 Results

In this section, we describe the regression technique and motivate the chosen performance models (cf. section 3.1). We report the overall findings of the regression analysis, in particular, the absent expertise effects (cf. section 3.2), the trial-number effects (cf. section 3.3), and finally we infer the underlying diagrammatic reasoning of users from our results (cf. section 3.4).

3.1 Regression Approach and Performance Models

Multiple regression techniques allow us to assess multiple correlations of (both numerical and categorical) explanatory (or independent) variables with a specific (dependent) performance variable [19]. In other words, a regression analysis makes it possible to determine whether there are effects (so-called significant predictors) of each explanatory variable which dominate the other explanatory variables included in the regression model. In particular, we apply linear mixed-effects modelling, which allows us to model dependencies in the observations as the answers of each participant are not

considered independent. This means that we may infer expertise effects, trial-number effects, and potentially the underlying diagrammatic reasoning.

Our choice of dependent variables were guided by indicators reflecting the characteristics of expert performance. Following [20], we expect expert performance to be more correct, faster and reflecting deeper problem representation compared to non-expert performance. In our experiment, we develop three corresponding performance models with correctness, response time (defined as the sum of processing and answering time on each question.) and diagram-fixation time (defined as the sum of all fixations over 200 ms on the screen in the relevant AOI) as dependent variables, respectively. In practice, we apply a stepwise forward variable selection procedure, i.e. a bottom-up approach, in which we test variables one at a time ending with those explanatory variables most central to performance. Only significant variables were retained in the final models (cf. table A2).

3.2 Expertise Effects

Our three performance models were expected to reveal expertise effects from the subset of explanatory variables measuring expertise levels.

In the first regression model, **correctness** was the dependent variable, and trial number and response time as well as self-assessed performance on diagrams and assessment of information coverage were significant predictors (main effects), while no interactions between explanatory variables appeared (cf. the first column of table A2). In the second regression model, **response time on correct answers** was the dependent variable, and trial number in the block of questions concerning each target concept as well trial number in the entire experiment were significant (cf. the second column of table A2), but the latter interacts significantly with question type. In the final regression model, **diagram fixation time on correct answers** was the dependent variable, and diagram position as well as fixation time on questions and trial number were significant predictors (cf. the third column of table A2), the interaction with question type reappears. The graphical plots of the regression models are displayed in figure 2.

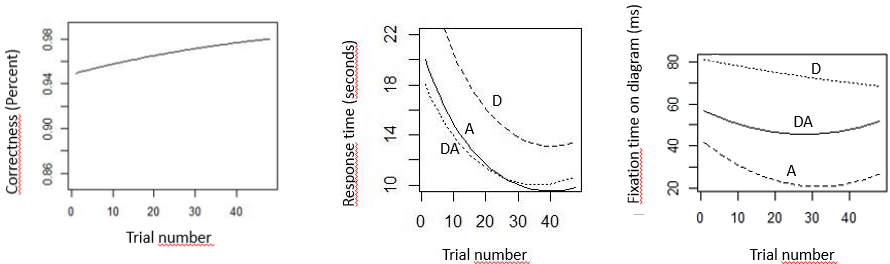


Fig. 2. Performance models with trial number of the experiment on the horizontal axes and dependent variables on the vertical axes. **To the left:** The vertical axis shows the probability of correctness (in percent), which slightly increases as the experiment proceeds. **In the middle:** The vertical axis shows the response time (in seconds) on correct observations on each question type (D, A and DA), which decreases as the experiment proceeds. **To the right:** The vertical

axis shows diagram-fixation time (in milliseconds) on correct observations on each question type (D, A and DA), which slightly decreases in the beginning of experiment.

It should be noted that the regression models showed no expertise effects (i.e. no explanatory variables reflecting expertise were significant predictors). This opens for several possible (opposite) interpretations: Expertise effects may be reduced due to the potential overload of users' limited processing capacity [21], absent due to insufficient expertise measures [22], or reversed due to information redundancy of the dual-entry modes [23].

3.3 Trial-Number Effects

In this section, we describe the significant trial-number effects, which appeared in all three performance models:

Correctness constituted our first performance indicator, and participants were able to understand the questions, retrieve answers from each of the entry modes and produce high overall correctness. It should be noted that question type was a non-significant predictor, which indicated that each of the question types were equally difficult. As the experiment proceeded, participants showed (slightly) increasing correctness (i.e. positive trial-number effect in the correctness analysis), cf. the left panel of figure 2.

Response time on correct answers constituted our second performance indicator (speed), and we see a significant interaction between performance and question type. In particular, D-questions required significantly longer response time compared to A- and DA-questions. As the experiment proceeded, and especially in the beginning of the experiment, the regression model showed that participants' response time on correct answers decreased implying increased performance (i.e. negative trial-number effect in the response-time analysis), cf. the middle panel of figure 2.

Diagram-fixation time on correct answers constituted our final performance indicator (depth). A significant interaction between performance and question type reappeared in the diagram-fixation-time model, where D-questions required significantly longer diagram-fixation time compared to A- and DA-questions, but now DA-questions required significantly longer diagram-fixation compared to A-questions. As the experiment proceeded, the regression model showed that participants' diagram-fixation time (slightly) decreased (i.e. weak negative trial-number effect in the diagram-fixation-time analysis), cf. the right panel of figure 2.

3.4 Diagrammatic Reasoning

In this section, we interpret the trial-number effects of the performance models to infer the underlying diagrammatic reasoning. Three implications for diagrammatic reasoning of terminology visualization can be inferred:

First, overall correctness is high, and we see that question type is non-significant. Hence, question types are fully comparable, and in particular, D-questions are not resulting in lower correctness, which implies that diagrammatic reasoning took place at the same level as “non-diagrammatic reasoning” of articles.

Second, speed on correct answers is increasing since response time is decreasing as the experiment proceeds for all three types of questions (cf. the profiles of the curves in the middle panel of figure 2) suggesting improved diagrammatic reasoning. However, performance on D-questions are lower as we see that D-questions are more (response) time consuming and require significantly longer response time compared to A-questions and DA-questions to produce correct answers (cf. the positions of the curves in the middle panel of figure 2). Hence, diagrammatic reasoning of the D-questions is the least efficient compared to A- and DA-questions.

Third, when we investigate diagram-fixation time, we see that on the DA-questions, diagram-fixation time is significantly longer than for A-questions but below D-questions (cf. the position of the curves in the right panel of figure 2). In particular, in the DA-questions we see that diagrams are fixated but that does not translate into longer total response time on correct answers. We suspect that answers are not retrieved from the diagram, because that would have required fixation time at the correspondingly high level as D-questions (since we might assume the difficulty of the three question types to be fully comparable). Instead, it is likely that users are confused by the diagram in the DA-questions before (quickly) retrieving the answer from the article and that points to the conclusion that diagrammatic reasoning is inefficient. Moreover, diagram fixation does not improve on the DA-questions as the experiment proceeds (cf. the horizontal profile of the DA-curve in the right panel of figure 2).

4 Conclusion

We may conclude that compared to the textual (article) entry mode, the graphical (diagram) entry mode is the most time-consuming mode for conveying domain-specific terminology and knowledge and only in the case of diagram-related questions (D-questions) are we able to provide evidence for effective and improving diagrammatic reasoning. In the case of DA-questions, users' diagrammatic reasoning is most likely inefficient and substituted by "non-diagrammatic reasoning" of the article. Therefore, we are not able to conclude whether dual access to both graphical and textual visualizations of domain-specific terminology potentially enhance target users' knowledge acquisition. It depends on the specific information need (question type). However, regarding the overall research question, the results imply that the visualization of concepts by means of terminological ontologies should be an integral part of in the interface of terminological resources as users are able to reason about and acquire knowledge from diagrams despite relatively long reasoning times.

We focused our investigation to response time and diagram fixation on questions, which participants answered correctly. It should be noted that long response time need not necessarily reflect long fixation time. Response time may also be a matter of long browsing time as participants travel a long search route across the dual-entry mode producing long scan paths (with few fixations). In particular, our experimental design containing multiple-choice questions placed above the stimulus-space (cf. figure 1), is prone to produce long scan paths as participants need to double-check retrieved answers with the available answers prior to the actual keying-in of the answer. Therefore, scan-paths should be included in future work on cognitive processing of dual-entry modes.

Acknowledgements. The study was funded by the VELUX FOUNDATION and constituted a sub-project of the DanTermBank project which aimed at developing the foundations for the establishment of a terminology and knowledge base in Denmark [24]. Thank you for all the support and encouragement to supervisors, team members and colleagues at IBC.

A Appendix

Table A1. Question types, questions and available answers of the target concept “energy tax” (*energiskat*) translated into English.

Question type	Question	Available answers
D1	How many types of energy taxes exist?	1: Four; 2: Six; 3: Eight
D2	What separates carbondioxide tax from duty on nitrogen oxides?	1: Purpose; 2: Content; 3: Taxpayer
A1	What can 'energy tax' be translated into in Danish?	1 (<i>energiavgift</i>): 'energy duty' ; 2 (<i>energiskat</i>): 'energy tax'; 3 (<i>energitakst</i>): 'energy rate'
A2	Energy taxes constituted 44 per cent of excise duties in 2011 according to whom?	1: OECD; 2: Eurostat; 3: Statistics Denmark
DA1	What type of tax or duty is energy tax?	1: Environmental duty; 2: Energy duty; 3: Excise duty
DA2	What is the purpose of energy tax?	1: Limiting environmentally damaging energy consumption; 2: Limiting environmentally damaging consumption; 3: Limiting environmental damage.

Table A2. Overview of regression models. Performance indicators (correctness, speed and depth) are dependent variable in each of the performance models. Explanatory variables are ordered by importance for explaining the variance in dependent variables beginning with the least important. “SIG” reflects a significant effect, while “NS” indicates nonsignificant effect, and “NA” indicates that the variable was irrelevant for that model. “Q” indicates a nonlinear effect, “POS” indicate a positive effect and “NEG” a negative effect.

	Correctness	Speed (Response time)	Depth (Diagram fixation)
Explanatory variable			
Display side of answer	NS	NS	SIG (NEG)
Total response time	SIG (NEG)	NA	NA
Self-rated search expertise	NS	NS	NS
Number of weekly Google search	NS	NS	NS
View on A mode	NS	NS	NS
View on D mode	NS	NS	NS
View on performance in A	NS	NS	NS
View on performance in D	SIG (POS)	NS	NS
Preference for D compared to A	NS	NS	NS
Preference for None compared to A	NS	NS	NS
View on information modes	SIG (POS)	NS	NS
Self-rated tax expertise	NS	NS	NS
Exposure to specialized texts	NS	NS	NS
Motivation	NS	NS	NS
Age	NS	NS	NS
Gender	NS	NS	NS
Question type D compared to A	NS	SIG (POS)	SIG (POS)
Question type DA compared to A	NS	NS	SIG (POS)
Trial number	SIG (POS)	SIG (Q)	SIG (Q)
Block trial number	NS	SIG (NEG)	NS

References

1. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C. & Giannopoulou, E.: Ontology visualization methods - a survey. *ACM Computing Surveys (CSUR)*, 39 (4), 10 (2007)
2. Polguère, A.: From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, ecu017 (2014)
3. Lew, R.: Multimodal lexicography: The representation of meaning in electronic dictionaries. *Lexikos*, 20(1), 290-306 (2010)
4. Seufert, T.: Supporting coherence formation in learning from multiple representations. *Learning and instruction*, 13(2), 227-237 (2003)
5. De Schryver, G. M.: Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography*, 16(2), 143-199 (2003)
6. Cabré, M. T.: Theories of terminology: Their description, prescription and explanation. *Terminology*, 9(2), 163-199 (2003)
7. Roche, C., Calberg-Challot, M., Damas, L., & Rouard, P.: Ontoterminology: A new paradigm for terminology. In *International Conference on Knowledge Engineering and Ontology Development* (pp. 321-326) (2009)
8. Madsen, B. N.: Terminological ontologies: Applications and principles. *Semantic Systems: From Visions to Applications. Proceedings of the Semantics 2006. Österreichische Computer Gesellschaft*, 271-282 (2006)
9. Madsen, B. N., Thomsen, H. E., & Vikner, C.: Comparison of Principles Applying to Domain-Specific versus General Ontologies. In *OntoLex* (pp. 90-95) (2004)
10. Guarino, N.: Formal ontology, conceptual analysis and knowledge representation. *International journal of human-computer studies*, 43(5), 625-640 (1995)
11. ISO: Terminology Work - Vocabulary. Part 1: Theory and application (1087-1). Geneva. (2000)
12. Ormerod, T. C., & Ball, L. J.: Cognitive psychology. *Handbook of qualitative research in psychology*, 554-574 (2008)
13. Nielsen, L. P.: Knowledge Dissemination Based on Terminological Ontologies. Using eye Tracking to Further user Interface Design (Doctoral dissertation, Department of International Business Communication, Copenhagen Business School (2015)
14. Tono, Y.: Application of eye-tracking in EFL learners' dictionary look-up process research. *International Journal of Lexicography*, 24(1), 124-153 (2011)
15. Bundesen, C. & Habekost, T.: Principles of visual attention: Linking mind and brain. Oxford University Press (2008)
16. Just, M. A., & Carpenter, P. A.: A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4), 329 (1980)
17. Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J.: Eye tracking: A comprehensive guide to methods and measures. Oxford University Press (2011)
18. Britton, B. K., Stimson, M., Stennett, B., & Gülgöz, S.: Learning from instructional text: Test of an individual-differences model. *Journal of Educational Psychology*, 90(3), 476 (1998)
19. Balling, L. W.: Morphological Effects in Danish Auditory Word Recognition (Doctoral dissertation, Aarhus Universitet Aarhus University, [Enhedsstruktur før 1.7. 2011] Aarhus University, Det Humanistiske Fakultet Faculty of Humanities, Afdeling for Engelsk Department of English) (2008)
20. Rikers, R. M., & Paas, F.: Recent advances in expertise research. *Applied Cognitive Psychology*, 19(2), 145-149 (2005)

21. Mayer, R. E., & Moreno, R.: Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, 38(1), 43-52 (2003)
22. Alexander, P. A.: Domain knowledge – evolving themes and emerging concerns. *Educational Psychologist*, 27 (1), 33-51 (1992)
23. Kalyuga, S., & Sweller, J.: Measuring Knowledge to Optimize Cognitive Load Factors During Instruction. *Journal of educational psychology*, 96(3), 558 (2004)
24. DanTermBank, <http://dantermbank.cbs.dk>

Towards Concept Maps 3.0: Visual Learning Designs as Web Data

Lars Johnsen¹, and Jesper Jensen¹

¹Department of Design and Communication, University of Southern Denmark, Kolding, Denmark
{larsjo, jesjen}@sdu.dk

Abstract. In this paper, it is proposed how concept maps may be described, annotated and exposed on the Web of Data, also known as Web 3.0. The paper briefly introduces concept maps as visual learning designs and goes on to describe three generations of web-based concept maps each reflecting different generations of web technology. The paper then defines the notion of concept maps 3.0 on the basis of five fundamental requirements. Finally, it is exemplified how concept maps 3.0 may be semantically marked up using the vocabularies schema.org and CXL and the data format JSON-LD.

Keywords: Concept maps 3.0, metadata, web data principles, schema.org, CXL

1 Introduction

The main goal of this paper is to propose how concept maps may be described, annotated and exposed on the Web of Data, also known as Web 3.0. The proposed changes to how concept maps should be represented on Web 3.0 do not, however, change what a concept map fundamentally is. At its core, a concept map is a graphical tool or visual representation, which can be used to express personal knowledge in a way that is easily understood by others. A concept map revolves round answering a focus question, and is often hierarchical in the sense that it contains a root node/concept, which represents the main topic of the map. Furthermore, a concept map comprises concepts (or instances of these) and linking phrases forming propositions that describe the relationship between concepts. When creating a concept map, there are no restrictions on which words or phrases, or what visual signals (shapes, color, etc.), one is allowed to use in order to represent one's knowledge of a topic [1].

From a learning perspective, concept maps have, among other things, been used as teaching material, as a way of supporting collaborative learning, or as a tool for evaluating student understanding of a specific topic because each concept map provides a (hopefully) clear visual representation of each student's personal understanding of the central topic [1]. By nature, concepts are visual in the way they map out relationships between concepts. This visual nature is further enhanced by various concept mapping software that supports the inclusion of, or references to, content in other modalities such as images, sound, video or documents on a computer or on the web.

2 Concept Maps 1.0 to 3.0 – Form and Functionality

When addressing concept mapping in a web-based context we find it useful to differentiate between different generations of concept maps based on how they are published and on their form and functionality in general. The actual representation of concept maps on the web comprises aspects such as representation of source code and data, visual representation and interactivity. Naturally, these aspects of form and functionality are shaped by the possibilities and limitations of the web technologies that are available at any given point in time in the history of the web.

In the following, we make the distinction between concept maps 1.0, 2.0 and 3.0 as three separate generations of concept maps on the web – corresponding to webs 1.0, 2.0 and 3.0 respectively. These three generations of web technology have gradually presented new ways of facilitating web-based concept mapping and provided additional opportunities of expanding the notion of what a web-based concept map is and what it can do.

Concept maps 1.0 can be considered the most basic form of web-based concept maps. The technologies of web 1.0 allow for the creation of static concept maps, as exemplified by those one may create in an offline, desktop concept mapping application such as CmapTools (<http://cmap.ihmc.us/cmaptools/>). As such, any text, image, or other content within a concept map 1.0 is static in the sense that it is fixed in the source code of the web page on which the concept map appears, and will only change if changes are made directly in the source code.

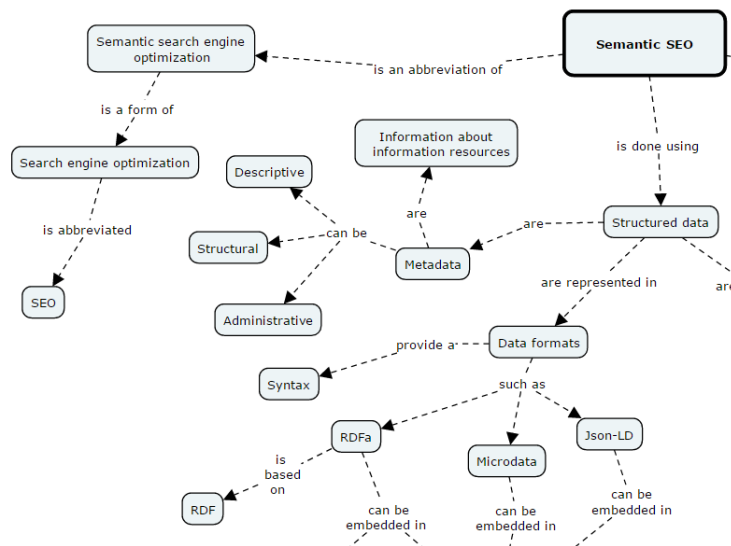


Fig. 1. Screenshot showing a part of a traditional concept map 1.0 about semantic search engine optimization created in an offline, desktop concept mapping application (CmapTools).

Concept maps 2.0 extend the accessibility and usability of concept maps 1.0. Supporting principles of the social web (web 2.0) such as collaborative content creation, sharing of content, and publishing content in open formats, tools like Cmap Cloud (<https://cmapcloud.ihmc.us/>) make it possible for users to produce and publish concept maps on the web. Much like its offline desktop predecessor CmapTools, Cmap Cloud allows users to attach resources such as documents, images, video, sound, and hyperlinks to web pages and other web-based content directly to individual concepts within a concept map. However, by offering its concept maps online in a dedicated environment that facilitates sharing and collaboration of these concept maps, Cmap Cloud brings concept mapping into the realm of web 2.0. In addition, links to other concept maps can be associated with concepts, thus allowing users, with whom a concept map has been shared, to browse not only that concept map itself, but also explore other concept maps and resources, which the concept map links to.

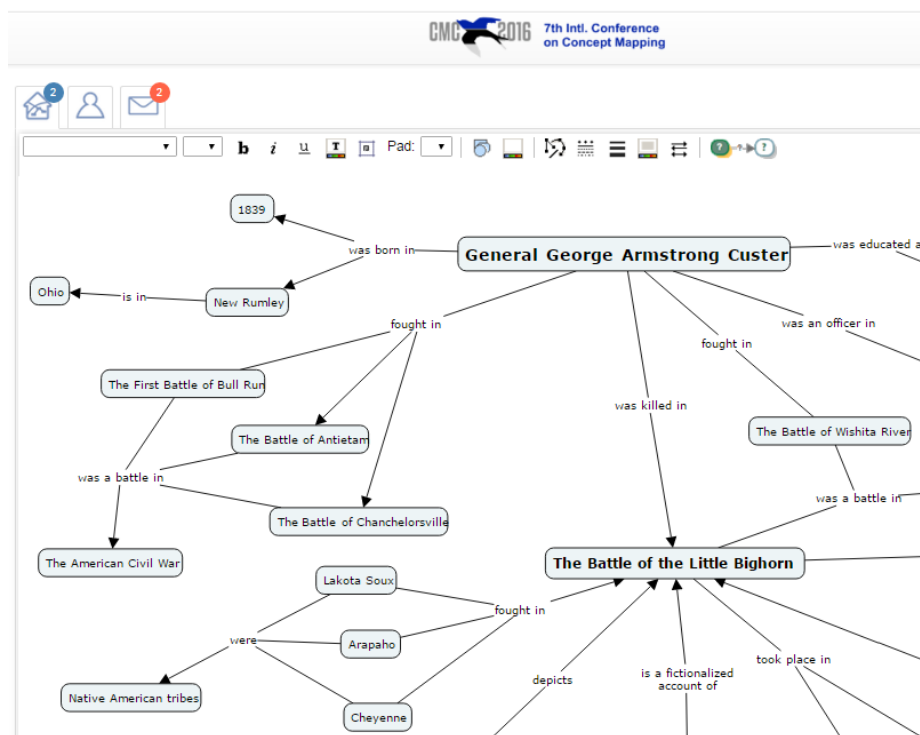


Fig. 2. Screenshot showing a part of a concept map 2.0 about the American general G.A. Custer created in the online concept mapping environment Cmap Cloud.

The idea of concept maps 3.0 involves the use of web 3.0 (semantic web) technologies in a concept mapping context. In other words, it is about making concept maps into machine-interpretable semantic web resources, and possibly even semantic learning resources, by integrating metadata into the source code of the concept maps. What makes concept maps 3.0 particularly interesting to explore is that they provide solu-

tions to key limitations found in concept maps 2.0. First, applying metadata, which provide detailed and machine-readable information about the different concepts and propositions of a concept map, can facilitate a higher level of discoverability, provided search engines such as Google are able to understand these metadata. This enables search engines to more easily discover concept maps, which are relevant to a specific search query. Second, one key feature of web 3.0 is that it supports the integration of data. Unlike previous generations of web technology, it is possible to identify the “meaning” of concepts with unique URL identifiers. The open data repository Wikidata (<https://www.wikidata.org>) is ideal for this, as it provides identifiers for all entities and concepts contained in it, as well as for the properties that describe them. These identifiers may be used in concept maps not only to uniquely identify concepts and concept types in the map but also to link directly into Wikidata. This in turn has the implication that Wikidata content can be automatically and dynamically identified, extracted and integrated with the original content of the concept map.

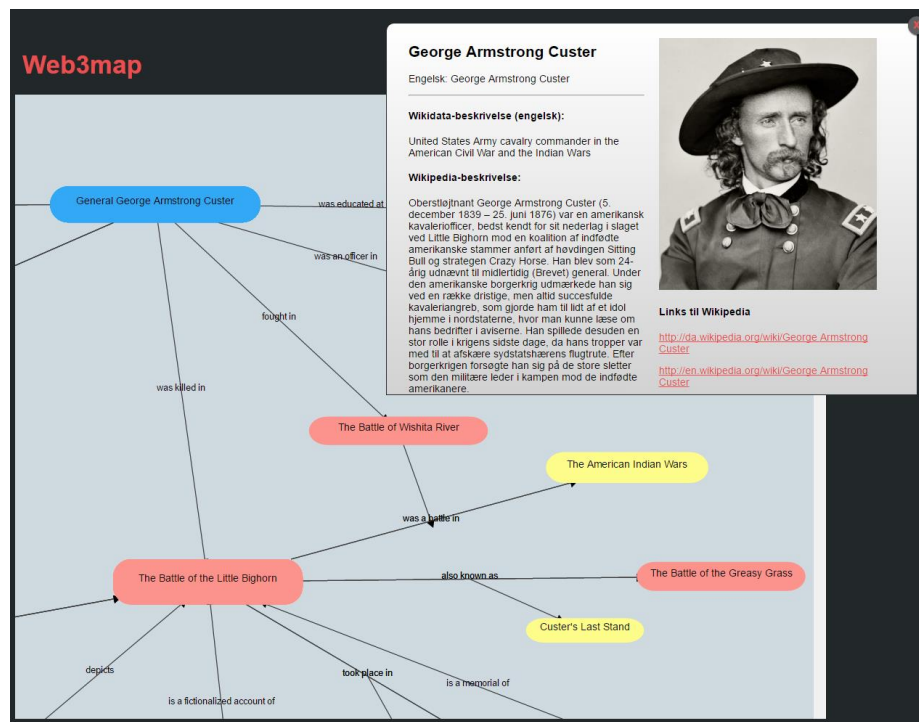


Fig. 3. This screenshot shows part of an early prototype of the web3map application – an example of a concept maps 3.0 application. When a user clicks on a concept in a concept map within the application, it automatically extracts and presents content from Wikipedia.org and Wikidata.org using Wikidata URL identifiers.

Another interesting possibility that comes with utilizing metadata and even integration of external semantic data is the ability to generate varied and dynamic visualiza-

tions of the data. Because the content of a concept map 3.0 is exposed in a specific semantic data format, it is possible to present its content, along with other external data, in ways that are vastly different from the visual presentation of the concept map itself. Content or data about places could be presented through the use of Google Maps. Data regarding historical events could be presented in timelines, for instance with tools offered by <http://histropedia.com/>. For people mentioned in a concept map 3.0 a visualization of a family tree including images and information about dates of birth and death might be generated.

3 Concept Maps 3.0 - Fundamental Requirements

Concept maps 3.0 are, as defined in the present context, learning resources exposed on the Web of Data. This means that they are not only published on the web as verbo-visuals to be accessed and interpreted by humans but also, at least in part, as sets of data to be discovered and consumed by software of different kinds.

The question is, then, how do we go about mapping semiotic structures like concept maps into machine-readable representations that can be hooked into, as it were, the Web of Data? What principles should be adhered to and what technologies applied?

Quite recently, The World Wide Web Consortium (W3C) has made available an extensive set of "best practices" for publishing data on the web [5]. These recommendations are no doubt valuable for publishers of (complex) web data on a substantial scale but arguably overkill as far as the publication of concept maps 3.0 is concerned. Here the more succinct *Web Data Principles* [8] seem more readily applicable (<http://dret.github.io/webdata/>). The Web Data Principles are defined as "a simple set of guidelines about how to make structured information more useful on the web" and consist of five recommendations of what should characterize data sets on the web and their distributions:

- Linkable
- Parseable
- Understandable
- Linked
- Usable

Adopting these recommendations, we define the following fundamental requirements for concept maps 3.0 as data sets:

- Concept maps should be linkable, that is accessible via persistent or stable identifiers. This obviously applies to the concept map as a whole but preferably also to its constituent parts. In this way, external resources can point to specific entities or objects in the structure.
- Concept map distributions should be represented in open formats that do not require proprietary software for processing and whose source code is open to inspection.

- Concept maps should be annotated by metadata using "well-known" and/or "well-documented" vocabularies.
- Concept maps should be linked to other resources to enhance their informational or learning value. Links should be typed if possible to signal their communicational purpose and/or the nature of their target and to enable automatic processing. Individual concepts should be linked to external resources to better determine their identity.
- Concept maps should be labeled with a license to signify when, where, how and by whom they may be put to use and under what circumstances.

For a concept map, and its constituent parts, to be *linkable*, it needs to be encoded in a format that allows identifiers to be attached to individual elements like concepts and propositions and groupings of these. This is not easily doable in traditional image formats where images usually are stored as unanalyzable wholes or "blobs". Instead, we propose that concept maps be represented in Scalable Vector Graphics (SVG), an XML (Extensible Markup Language) language for (animated) two dimensional graphics, and that unique identifiers be attached to all verbo-visual elements that manifest concept map constructs (focus question, concepts, propositions, etc.). To do so, visual elements like `<circle>` and `<line>` need to be grouped with `<text>` elements using the `<g>` element. For instance:

```
<svg width="100" height="100">
  <g id="globalwarming">
    <desc>Concept</desc>
    <circle cx="50" cy="50" r="40" stroke-width="4"
fill="gray" />
    <text fill="white" x="25" y="50">Global warm-
ing</text>
  </g>
</svg>
```

Here, the concept of global warming is symbolized by a group of elements: a gray circle, the string "Global warming" and a piece of descriptive non-displayable metadata (desc). The group is uniquely identified and embedded in a 100 x 100 SVG canvas. X's and y's indicate positions in the canvas, and r the radius of the circle.

Employing SVG as the primary encoding format, the second requirement of being *parseable* is automatically fulfilled. SVG is an open format as well as an open standard endorsed by the W3C. A further benefit of utilizing SVG for storing concept maps is that SVG is supported by a great deal of user agents, most notably browsers. Finally, SVG code can be directly embedded in HTML (HyperText Markup Language) and thus rendered as part of a larger web page.

To be *understandable*, and hence discoverable and conducive to processing, concept maps must be described and annotated using a "well-documented" vocabulary and preferably one that is accessible on the Web. One such language for specifying

concept map structure does exist, namely the Concept Mapping Extensible Language (<http://cmap.ihmc.us/xml/CXL.html>). As the name indicates, it is really the underlying schema for an XML-based language, CXL, for marking up concept maps and their contents. The problem is that this vocabulary is not very "well-known" outside the concept maps community and not widely supported by non-concept maps software. For instance, it is not known by major search engines such as Google, Bing and Yandex, which primarily support schema.org (<https://schema.org/>). Schema.org is a general vocabulary for labeling things search engines "care about" - persons, places, products, events and various sorts of creative works (books, films, apps and so forth). Not surprisingly, schema.org does not contain categories and properties to capture notions intrinsic to concept maps (focus question, proposition, etc.). It is, however, possible to include references to externally defined types in schema.org. We therefore propose to use schema.org as the main vocabulary to mark up concept maps and their contents but recommend the use of the CXL model as an ancillary mechanism to identify concept maps terms.

Adding schema.org/CXL metadata to SVG concept maps can be done in various ways. One method is to use the format RDFa (Resource Description Framework in Attributes), which lets concept mappers embed annotations in the SVG code directly. In this way, metadata travel, so to speak, with the concept map they describe. It is also possible, however, to detach semantic metadata from the concept map itself and store them separately. There are several advantages in doing so. One is that the SVG code is kept clean and is easier to read. The other is that metadata can be added to concept maps without one having direct writing rights to these. Annotating concept maps via reference rather than embedding requires a format like JSON-LD (<http://json-ld.org/>), a relative newcomer to the field of semantic markup. JSON-LD is an advanced form of JSON (JavaScript Object Notation), which allows for the inclusion of vocabularies. A further potential of JSON-LD is that it may be stored in a separate file or as an integral part of an HTML document. An example of schema.org/CXL metadata in JSON-LD is given and explained below.

To make concept maps *linked* is tantamount to providing their users with access to additional content on the Web thus encouraging more explorative learning activities or deeper learning of specific topics. But to software, links can also act as identifiers, i.e. as pointers referencing web pages, which unambiguously indicate the meaning or identity of some concept. For example, a link to the Wikidata page <https://www.wikidata.org/wiki/Q19643> in some concept map would indicate that the concept of "Queen" is to be construed as a female monarch in that concept map and not as a, say, chess piece, a playing card or the famous British rock band. (And as a bonus, that Wikidata page would itself provide additional structured information, which might be extracted and integrated into the concept map).

The last requirement of concept maps 3.0 is that they be labeled with information about how *usable* they are. Linking to a Creative Commons license is one appropriate solution to this requirement.

4 Representing Concept Maps 3.0

It is beyond the scope of this paper to give a full account of possibilities and constraints of marking up concept maps in schema.org/CXL using JSON-LD. A snippet of code specifying metadata for a history concept map about the American general George Armstrong Custer, however, hints at what such a solution might entail:

```
<script type="application/ld+json">
{
  "@context": "http://schema.org/",
  "@type": "CreativeWork",
  "learningResourceType" : "concept map",
  "inLanguage" : "en",
  "additionalType" :
"http://cmap.ihmc.us/xml/CXL.html#cmap",
  "name": "Custer",
  "url":
"https://cmapscloud.ihmc.us/viewer/cmap/1PXQ8ZZHR-
22371RZ-16M4BB",
  "description/focusQuestion": "What was General George
Armstrong Custer famous for?",
  "potentialAction" :
  {
    "@type": "SearchAction",
    "query":
"https://cse.google.dk/cse/publicurl?cx=01527297755418971
4981:bisamwwcwbe&q={concept name}",
    "description" : "Search for Books about the concept
map's concepts using a Google Custom Search Engine",
    "result" : "List of books associated with the se-
lected concept"
  },
  "mainEntity" :
  { "@type" : "Person",
    "additionalType" :
"http://cmap.ihmc.us/xml/CXL.html#concept",
    "sameAs" :
"https://www.wikidata.org/wiki/Q188205",
    "url":
"https://cmapscloud.ihmc.us/viewer/cmap/1PXQ8ZZHR-
22371RZ-16M4BB#custer",
    "name" : "George Armstrong Custer",
```



```

        "image" :
        "https://upload.wikimedia.org/wikipedia/commons/1/16/Custer_Bvt_MG_Geo_A_1865_LC-BH831-365-crop.jpg",
        "description" :
        {
            "@type" : "Role",
            "roleName" : "Google's Knowledge Graph",
            "description" : "http://g.co/kg/m/0pzgm"
        }
    },
    "about" :
    { "@type" : "Event",
      "additionalType" :
      "http://cmap.ihmc.us/xml/CXL.html#concept",
      "sameAs" : "https://www.wikidata.org/wiki/Q205422",
      "name" : "The Battle of The Little Bighorn",
      "alternateName" : "Custer's Last Stand"}
  }
</script>

```

This example demonstrates how selected types of metadata can be applied to a concept map. Administrative metadata such as "learningResourceType" and "inLanguage" characterize this map as a whole while descriptive metadata are employed to provide information about the individual entities mentioned in the map. Thus, it is indicated that the main entity of the concept map is a person named "George Armstrong Custer" to which a description and an image are attached. A so-called fragment identifier (#custer) is supplied to point to the actual verbo-visual representation of Custer in the concept map. Further, it is stated that the concept map is also about an "event" with a given and an alternate name. Both entities have links that reference pages in Wikidata signifying their identity. Last but not least, the concept map is connected to a so-called (potential) Action specifying conditions under which the user may search for relevant books on the Web about selected entities in the concept map. In other words, the metadata may not only provide information about the concept map and its contents but also about how it may be acted upon by the user and what results these actions may yield. The example also illustrates a couple of nifty mechanisms in the schema.org vocabulary for extending categories and properties:

- The "additionalType" property lets the concept mapper refer to terms in the CXL schema. In this way, a CreativeWork in schema.org can at the same time be a concept map in CXL and Custer a person in schema.org and a concept in CXL.
- The slash (/) convention facilitates the restriction of properties. For instance, schema.org does not have a property or type to specify the focus question of a concept map. This can be done, however, by restricting the "description" property.
- The "Role" type permits properties to be renamed and reinterpreted. This means, for example, that a simple description property can be turned into "Google

Knowledge Graph" to indicate Google Knowledge Graph's description of some entity.

[Hent webadresse](#)
[Eksempler](#)
[JSON-LD](#)

VALIDER

Logoer (hvad er det?)

```

1 <script type="application/ld+json">
2 {
3   "@context": "http://schema.org/",
4   "@type": "CreativeWork",
5   "learningResourceType" : "concept map",
6   "inLanguage" : "en",
7   "additionalType" : "http://cmap.ihmc.us/xml/CXL.html#cmap",
8   "name": "Custer",
9   "url": "https://cmapscloud.ihmc.us/viewer/cmap/1PXQ8ZZHR-
10 22371RZ-16M4BB",
11   "description/focusQuestion": "What was General George
12 Armstrong Custer famous for?",
13   "potentialAction" :
14   {
15     "@type": "SearchAction",
16     "query": "https://cse.google.dk/cse/publicurl?
17 cx=015272977554189714981:bisamwwcwb&q={concept name}",
18     "description" : "Search for Books about the concept
19 map's concepts using a Google Custom Search Engine",
20     "result" : "List of books associated with the selected

```

Resultater - [Udvalgte uddrag for begivenheder](#) (hvad er det?) 2 Fejl

CreativeWork (1) 2 Fejl

CreativeWork
learningResourceType: concept map
inLanguage: en
additionalType: http://cmap.ihmc.us/xml/CXL.html#cmap
name: Custer
url: https://cmapscloud.ihmc.us/viewer/cmap/1PXQ8ZZHR-22371RZ-16M4BB
description/focusQuestion: What was General George Armstrong Custer famous for?
potentialAction [SearchAction]:
query: https://cse.google.dk/cse/publicurl?cx=015272977554189714981:bisamwwcwb&q={concept name}
description: Search for Books about the concept map's concepts using a Google Custom Search Engine
result [Thing]:
name: List of books associated with the selected con

Fig. 4. This screenshot displays how Google's Structured Data Testing Tool reads and validates the metadata for the Custer concept map.

References

1. Cañas, A.J., Carff R., Hill G., Carvalho, M., Arguedas, M., Eskridge, T.C, Lott J., & Carvajal, R. (2005): Concept Maps: Integrating Knowledge and Information Visualization. I Tergan, S.O. & Keller, T.: Knowledge and Information Visualization: Searching for Synergies. Heidelberg / New York: Springer Lecture Notes in Computer Science.
2. Eppler, M.J. (2006) A Comparison between Concept Maps, Mind Maps, Conceptual Diagrams, and Visual Metaphors as Complementary Tools for Knowledge Construction and Sharing, in: Information Visualization, 5(3): 202-210.
3. Jensen, J. (2016). Web 3.0's didaktiske potentialer. In N. B. Dohn & J. J. Hansen (Eds.), Didaktik, design og digitalisering (pp. 91-112). Frederiksberg: Samfundslitteratur.
4. Johnsen, L. (2016). Læringsobjekter 3.0. In N. B. Dohn & J. J. Hansen (Eds.), Didaktik, design og digitalisering (pp. 113-130). Frederiksberg: Samfundslitteratur.
5. Lóscio, B.F., Burle, C. & Calegari, N. (2016): Data on the Web Best Practices. W3C Working Draft 12 January 2016. <https://www.w3.org/TR/dwbp/>
6. Nesbit, J.C & Adescope, O.O (2013): Concept Maps for Learning: Theory, Research and Design. I Schraw, G., McCrudden, M.T. og Robinson, D.: Learning Through Visual Displays. Information Age Publishing.
7. Novak, J.D & Canas, A.J. (2006/2008): The Theory Underlying Concept Maps and How to Construct and Use Them. <http://cmap.ihmc.us/publications/researchpapers/theorycmaps/theoryunderlyingconceptmaps.htm>
8. Wilde, E. (2016): Web Data. Overview of the Web Data Principles. 17 Feb 2016. <http://dret.github.io/webdata/>