Effects of Surprisal and Locality on Danish Sentence Processing
An Eye-tracking Investigation

Balling, Laura Winther; Kizach, Johannes

Document Version
Accepted author manuscript

Published in:
Journal of Psycholinguistic Research

DOI:
10.1007/s10936-017-9482-2

Publication date:
2017

License
Unspecified

Citation for published version (APA):
Balling, L. W., & Kizach, J. (2017). Effects of Surprisal and Locality on Danish Sentence Processing: An Eye-tracking Investigation. *Journal of Psycholinguistic Research*, *46*(5), 1119-1136. https://doi.org/10.1007/s10936-017-9482-2

Effects of Surprisal and Locality on Danish Sentence Processing: An Eye-tracking Investigation

Laura Winther Balling and Johannes Kirzach

Journal article (Accepted manuscript)

CITE: Balling, L. W., & Kizach, J. (2017). Effects of Surprisal and Locality on Danish Sentence Processing: An Eye-tracking Investigation. Journal of Psycholinguistic Research, 46(5), 1119-1136. DOI: 10.1007/s10936-017-9482-2

EFFECTS OF SURPRISAL AND LOCALITY ON DANISH SENTENCE PROCESSING: AN EYE-TRACKING INVESTIGATION

Abstract

An eye-tracking experiment in Danish investigates two dominant accounts of sentence processing: locality-based theories that predict a processing advantage for sentences where the distance between the major syntactic heads is minimized, and the surprisal theory which predicts that processing time increases with big changes in the relative entropy of possible parses, sometimes leading to anti-locality effects. We consider both lexicalised surprisal, expressed in conditional trigram probabilities, and syntactic surprisal expressed in the manipulation of the expectedness of the second NP in Danish constructions with two postverbal NP-objects in Danish. An eye-tracking experiment showed a clear advantage for local syntactic relations, with only a marginal effect of lexicalised surprisal and no effect of syntactic surprisal. We conclude that surprisal has a relatively marginal effect, which may be clearest for verbs in verb-final languages, while locality is a robust predictor of sentence processing.

Keywords: Sentence processing; Eye tracking; Locality; Surprisal theory; Danish language

1. Introduction

Among many accounts of sentence processing, two broad classes currently stand out: those based on memory constraints and those based on expectation. Memory-based accounts argue that processing is limited by memory constraints, making local syntactic relations preferable to non-local relations (Gibson, 1998, 2000, 2003, Hawkins, 1990, 1994, 2004, 2014). Consider the dative-alternation where the goal argument can be realized as an NP (the NP-construction), as in (1), or as a PP (the PP-construction) as in (1):

(1) [**Peter**]$_{S}$ [**gav**]$_{V}$ [**drengen**]$_{IO}$ [**en** kage med flødeskum]$_{DO}$. NP-construction

[Peter]$_{S}$ *[gave]*$_{V}$ *[boy-the]*$_{IO}$ *[a cake with whipped cream]*$_{DO}$.

(2) [**Peter**]$_{S}$ [**gav**]$_{V}$ [**en** kage med flødeskum]$_{DO}$ [**til** drengen]$_{\sim IO}$. PP-construction

[Peter]$_{S}$ *[gave]*$_{V}$ *[a cake with whipped cream]*$_{DO}$ *[to boy-the]* $_{\sim IO}$.

Here, Hawkins' locality-based account would predict a preference for (1) over (2) because the syntactic heads of the constituents, marked with boldface above, are closer together in (1) than in (2), making the relations more local. Note that we assume that the parser will attempt to satisfy the subcategorization requirements of the ditransitive verb *give* 'give' as quickly as possible, which means that in (2) as soon as *til* 'to' is encountered the parser will attach the prepositional phrase as the goal-argument. Extensive arguments in favour of this assumption can be found in the literature and we will not discuss this further (see Fodor & Inoue, 1998, 2000; Frazier, 1987; Frazier & Clifton, 1996; Pritchett, 1992).

Gibson's (2000) locality-based account also predicts a preference for (1) over (2), though for slightly different reasons. He argues that dependencies, such as the ones between the verb and its

arguments, become more difficult when more referential expressions intervene. Referential expressions are defined as nouns and lexical verbs in Gibson's theory (Gibson, 2000, pp. 105–107). In (2), two referential expressions intervene between the verb *gav* 'gave' and the goal-argument *til drengen* 'to the boy', whereas in (1) only one referential expression intervenes between the verb and the second argument, resulting in a predicted preference for (1) over (2).

For head-initial languages like English and Danish, both Hawkins' and Gibson's theories predict a general preference for ordering short constituents before long ones, and this preference shows up in both production (De Cuypere & Verbeke, 2013; Hawkins, 1994, 1998, 2011, Kizach, 2012, 2014, 2015; Kizach & Vikner, 2016; Rosenbach, 2005; Seoane, 2009; Wasow, 2002; Wiechmann & Lohmann, 2013) and comprehension (Christensen, Kizach, & Nyvad, 2013; Gibson, 2003; Hofmeister & Sag, 2010; Kizach & Balling, 2013; Warren & Gibson, 2002).

In contrast, expectation-based accounts argue that processing difficulty varies with the expectedness of words and constituents. One way of construing expectedness is through the information theoretical concept of surprisal: the surprisal theory of sentence processing as formulated by Levy (2008) based on Hale (2001) posits that processing speed for word *n* will vary as a function of the amount of change between the probability distribution of possible sentence parses at word *n*-1 and the probability distribution at word *n*, measured through their relative entropy. The change in relative entropy is equal to the surprisal of word *n*. If word *n* has a low surprisal value, it should be relatively easy to process, while a high surprisal value should make it more difficult to process.

This theory can explain the anti-locality effects that have been reported in the sentence processing literature (Konieczny, 2000; Konieczny & Döring, 2003; Levy, 2008; Levy, Fedorenko, & Gibson, 2013; Levy & Keller, 2013; Smith & Levy, 2013; Vasishth & Lewis, 2006) alongside

the studies showing locality effects quoted above. To illustrate the point, we consider the following German example from Konieczny (2000, p. 631 his (1a) and (1b)):

(3) [**Er**]$_{S}$ [hat]$_{AUX}$ [das **Buch,** das Lisa gestern gekauft hatte,]$_{DO}$ [**hingelegt**]$_{MV}$

[He]$_{S}$ [has]$_{AUX}$ [the book that Lisa yesterday bought had,]$_{DO}$ [laid-down]$_{MV}$

(4) [**Er**]$_{S}$ [hat]$_{AUX}$ [das **Buch**]$_{DO\text{-}}$ [**hingelegt**]$_{MV}$, [das Lisa gestern gekauft hatte]$_{\text{-}DO}$

[He]$_{S}$ [has]$_{AUX}$ [the book]$_{DO\text{-}}$ [laid-down]$_{MV}$ [that Lisa yesterday bought had]$_{\text{-}DO}$

The verb *hingelegt* ('laid-down') is processed faster in (3) than in (4), even though the head noun in the direct object (*Buch* 'book') and the verb are further apart in (3) – hence the term anti-locality. In the examples in (3) and (4), a verb is already likely to occur following a subject, and this likelihood increases when more non-verb constituents occur after the subject. The longer the relative clause is, the more likely the verb becomes, and therefore the reading time for the verb decreases with increased relative clause length.

Both locality and surprisal may be indexed in different ways. For locality, Hawkins (1994, 2004) suggested measuring distance as the number of words, using the number of words intervening between the verb and the first element in the object as an indicator of the locality of the relation between these two constituents. If we consider the examples in (1) and (2) as an illustration, in (1) the verb is immediately followed by the indirect object *drengen* 'the boy' which is in turn followed by *en* ' a' – the first word of the direct object. This means that the heads of the three syntactic constituents in the VP are adjacent, which is the optimal situation from a locality perspective. In (2), on the other hand, the heads are not adjacent since the longer direct object intervenes between the verb and the shorter indirect object. From a locality point of view, the order in (2) is therefore predicted to be harder to process. Here, we use length difference in words as a convenient indicator

of locality, following Hawkins (1994, 2004, 2014). Various other measurements (nodes, syllables, letters) have been demonstrated to be highly correlated with this (Szmrecsanyi, 2004; Wasow, 1997, 2002). Gibson's (1998, 2000) discourse referent measure, which counts only referring expressions, makes the same prediction as Hawkins' word count measure regarding (1) and (2) as discussed above (cf. Hawkins, 2014, p. 56).

Interestingly, the locality metrics (whether they are word count or discourse referent count) do not change depending on which syntactic structure is posited for the NP-construction. Since the measured distance is between lexical items (the words that constitute the syntactic heads of the relevant phrases) it makes no difference if we assume a structure with a flat VP with three daughters (the verb and both arguments) or if we assume a structure with VP-shells. The only syntactic assumption we need to make is that there are two arguments and that each has a syntactic head.

Like locality, surprisal may be measured in different ways; in fact, the different ways of estimating surprisal arguably vary more than those of locality. The formulations of Hale (2001) and Levy (2008) are based on possible parses in probabilistic context-free grammars (PCFG) which are syntactic rather than lexicalised in the sense that the probabilities are based on parse trees of syntactic categories without any lexical information. Demberg and Keller (2008), by contrast, include both syntactic and lexicalised surprisal in their analyses of the Dundee corpus of eye-movements, finding effects of both syntactic surprisal based on PCFG parse tress and of forward transitional probabilities based on *n*-gram language models, which they see as a form of lexicalised surprisal. Frank and Bod (2011) compare hierarchical and linear language models that index surprisal in different ways and find that the hierarchical models do not explain variance in sentence processing data over and above what the linear models explain.

Here we compare different versions of surprisal pitted against locality, for a language that, to our knowledge, has not previously been investigated from this perspective, namely Danish. In an

eye-tracking experiment, we use NP-constructions of the type in (1a), with a factorial manipulation of the relation between NP1 and NP2 in terms of the order (short NP before long NP vs. long before short) and the length difference between the two NPs (differences of two, four or more words), as shown in table 1. The prediction from a locality point of view is that short-before-long orders are globally easier to process than long-before-short orders, but with regard to the length difference variable, the word count metric and the discourse referent metric make slightly different predictions. The word count metric predicts that the processing advantage should increase with increased length difference. The discourse referent metric, however, divides our three length differences into two groups predicting a processing advantage for examples with two or four word differences over examples with a difference of more than four words. The reason is that the difference in our materials between the two and four word examples is the presence or absence of prenominal adjectival modifiers. Since adjectives do not introduce new discourse referents in Gibson's (2000) theory, no difference in processing complexity is predicted.

[INSERT TABLE 1 APPROXIMATELY HERE]

Surprisal, by contrast, makes a prediction specifically for the second NP: the more material has already occurred in the sentence before NP2, i.e. the longer NP1 in the long-before-short orders, the more expected NP2 becomes. We note here that the direct object (NP2 in the NP-construction) is obligatory in Danish when the indirect object (NP1) is present. Under a surprisal account of sentence processing, this should make NP2 faster to process with increasing length of NP1.

Since a corpus of PCFG parse trees is not available for Danish, and not easily constructable, we model our manipulation on Konieczny's (2000) stimuli that showed anti-locality effects and are used by Levy (2008) as evidence for his surprisal theory. This surprisal manipulation at the onset of

NP2 as a consequence of different lengths of NP1 is clearly syntactic rather than lexicalised. We compare this conceptualisation of surprisal with linear lexicalised surprisal as expressed in conditional trigram probabilities of the words in the relevant constituents, following Demberg and Keller (2008).

2. Method

2.1 Design and materials

We constructed sets of six stimulus sentences: each set consisted of the same basic NP-construction, with the same verb and fundamentally the same referents, but manipulated so that we had two different orders of the NPs, short-before-long vs. long-before-short, and three different differences in length between the two NPs. Previous studies have detected a preference for definites to precede indefinites in the NP-construction in both Danish and English (Brown, Savova, & Gibson, 2012; Clifton & Frazier, 2004; Kizach & Balling, 2013), and others have demonstrated that definite NPs are less likely to be modified than indefinite ones (Thornton, MacDonald, & Arnold, 2000). We therefore kept all NPs indefinite to avoid confounding effects of this. All target clauses were followed by a coordinated main clause, to reduce wrap-up effects. An example set of sentences is shown in table 1.

The NPs were modified with premodifying adjectival phrases and post-modifying prepositional phrases and relative clauses. The sentences were constructed so that the same modifiers could be used for both NP1 and NP2, keeping the lexical variation at a minimum. To make sure that the modifiers were in fact equally compatible with the inanimate theme argument and the animate goal argument, we normed the stimuli. The same modifiers were used with an animate and an inanimate noun as in the examples in (5) and (6) below, where *æble* 'apple' and *assistent* 'assistant' are modified in the same way. We used the modified NPs in simple declarative

clauses distributed on two lists, so that no participant saw the same item in both the animate and the inanimate conditions.

(5) Arnold havde en helt vidunderlig assistent fra Danmark der var meget stor.
Arnold had a completely wonderful assistant from Denmark that was very big
'Arnold had a completely wonderful assistant from Denmark who was very big'

(6) Tyskeren spiste et helt vidunderligt æble fra Danmark der var meget stort.
German.the ate a completely wonderful apple from Denmark that was very big
'The German ate a completely wonderful apple from Denmark which was very big'

The sentences were presented using Google Drive to 38 participants, who judged the acceptability of each sentence on a scale from 1 to 5. The were no significant differences in the acceptability ratings between the animate and inanimate conditions (3.35 for the animate, and 3.58 for the inanimate, paired t-test result: $t(14) = -1.50$, $p = 0.16$).

We originally constructed 16 sets of six sentences, but one set was removed from the analyses because several participants remarked on its oddity and it also received the lowest ratings in the norming study. All in all, there were therefore 96 stimulus sentences, of which 90 were included in the analyses reported below. The sentences were distributed over two lists, so that each participant saw three sentences containing the same referents, and consequently half of the stimulus sentences. Reading three sentences with the same referents may induce some priming, but we control this by including a variable in our analysis that indexes for each sentence how many times a member of the same sentence set has been seen earlier in the experiment. In addition to the 48 stimulus sentences on each list, the experiment included twenty filler sentences, of varying degrees

of complexity, and three training items. Two-choice forced choice comprehension questions were included after ten of the filler sentences, to ensure that participants read for comprehension. See the appendix for a full list of stimulus sentences. The decision to repeat stimulus sentences was made for the practical purpose of keeping the experiment relatively short, making it feasible to run the experiment with volunteers within a reasonable time frame. This approach was possible because the experiment was not a decision task, where repetition would be more problematic, and because we could control the effect of repetition by including repetition as a variable in the analysis. In the analysis, the repetition variable turned out to show a rather interesting pattern, as we will return to below.

In addition to the length difference and order variables that together index syntactic surprisal, we also investigated the role of conditional trigram probability as a measure of lexicalised surprisal. The conditional trigram probability of a word expresses the probability that a target word will occur given that the two previous words have already occurred. For the word *flødeskum* 'whipped cream' in the context of the trigram *kage med flødeskum* 'cake with whipped cream', the calculation is the following:

$$p(flødeskum|kage\ med) = \frac{p(kage\ med\ flødeskum)}{p(kage\ med)}$$

The probability of *flødeskum* following *kage med* is estimated as the probability of the whole trigram *kage med flødeskum* divided by the probability of the initial bigram *kage med*; in other words, out of all the occurrences of *kage med* something, how often that something is *flødeskum*. The idea of using conditional *n*-gram probabilities as indicators of word predictability in context is based on MacDonald & Shillcock (2003), who showed that eye-movement measures were sensitive to the conditional bigram probability of a word. Here, we use conditional trigram

probability following Balling (2013) since trigram probabilities generally give higher accuracy than bigram probabilities (Koehn, 2010). We take this conditional probability as a measure that corresponds to lexicalised surprisal, based on the argument of Demberg and Keller (2008) that surprisal at "word w_{k+1} corresponds to the negative logarithm of the conditional probability of word w_{k+1} given the sentential context $w_1 \ldots w_k$". Here we restrict ourselves to calculating the probability of each word based on the two preceding words, rather than the entire preceding context; in practice, given the decreasing frequency of increasing *n*'s, these trigram probabilities are unlikely to be radically different from what probabilities based on higher *n*-grams would be.

To estimate conditional word trigram probabilities, we trained a language model on a large corpus of Danish text (the 56 million words of KorpusDK) using the SRI Language Modelling Toolkit (Stolcke, Zheng, Wang, & Abrash, 2011). We used modified Kneser-Ney smoothing (Chen & Goodman, 1999) to mitigate the problem of word trigrams in the experimental sentences being unattested in the corpus; the logic of this is that probabilities for unattested trigrams are estimated based on the probabilities of their constituent bigrams and the number of different contexts a word has appeared in (Jurafsky & Martin, 2009). Based on this language model, per word conditional trigram probabilities were extracted; these are log probabilities and were therefore summed across words in a sentence or constituent to give the total probability of this sentence or constituent. We use the log probabilities given by the model, rather than the negative log that constitutes surprisal, but the basic correspondence between the two posited by Demberg and Keller (2008) still holds.

2.2 Participants

Thirty-three participants were recruited among students at Copenhagen Business School; the data from three of these were discarded due to poor eye-tracking quality or low score on comprehension questions (below 80%). The 30 participants whose data were analysed were aged between 18 and

29 (mean 21.2), all had Danish as their native language and had normal or corrected-to-normal vision.

2.3 Task and procedure

The experiment was run on an Eyelink 1000 eye tracker with remote tracking of the right eye at a rate of 500 Hz. The participants were seated at a distance of approximately 60 cm from the screen. On arrival, the participants were instructed orally about the task, before being presented with written instructions on the screen. The eye tracker was then calibrated using a five-point grid. Following the calibration, participants read 71 sentences, of which ten were followed by comprehension questions; these occurred at random intervals. Each sentence occurred on a separate screen and participants pressed the space bar on the keyboard to move from one sentence to the next. After the eye-tracking experiment, brief background questions were asked, to establish age, native language(s) and vision. The entire procedure took about 15 minutes.

2.4 Statistical analyses

Our main explanatory variables in all analyses were the order of NPs (short-before-long vs. long-before-short) and the length difference between them (differences of two vs. four vs. more words); the combination of those two variables resulted in the six sentences per set of sentences exemplified in table 1 above. In addition, we considered the conditional trigram log probabilities, summed across the words of the sentence or constituent in question, as an index of lexicalised surprisal. We also included as control variables trial number (to index fatigue/learning effects), repetition of sentence set (encountering the basic referents and sentence scheme for the first, second or third time), and length in letters for a purely formal length variable, which is different from the syntactically oriented length difference variable.

All analyses were conducted in R (R Development Core Team, 2014) using linear mixed-effects models in the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) with p-values from the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2015). We applied a bottom-up analysis strategy, adding one predictor at a time and including the most control-oriented predictors before those relevant to the key questions. Non-significant predictors were discarded, except for the key explanatory variables length difference and order. We included random slopes and levels corresponding to the fixed-effects structure following Barr et al. (2013). As is often the case, the models with the most complex random effects structures frequently did not converge, in which case we used the most meaningful and complex random effects structure that did converge.

As part of our model criticism, we calculated variance inflation factors for all models to ensure that no effects were overly affected by collinearity (using the vif-function in the package car (Fox & Weisberg, 2011), adjusted for use with lmer-objects by Søren Feodor Nielsen). Further, the residuals of the models were inspected to investigate whether they were approximately normally distributed; when this was not the case, observations with large standardised residuals were excluded (removing between 1.6 and 2.1 percent of observations) and the models refitted. This procedure improved the distribution of the residuals without changing the conclusions. Apart from this model-based exclusion of outliers, we only excluded one observation, a case where the participant, presumably by accident, clicked past the sentence without reading it.

3. Results and discussion

Our analyses fall into two clusters: two more global analyses that investigate the predictions of locality accounts, and three more local analyses that compare the locality and surprisal accounts of sentence comprehension specifically on NP2.

3.1 Locality effects in global sentence comprehension

For our stimuli, a locality account of sentence processing would predict a general preference for short-before-long orders of NPs because this makes for more local syntactic relations, as illustrated in example 1 above. As can be seen from table 1, the syntactic relations in the long-before-short orders become increasingly non-local with increased length difference, and we would therefore expect the disadvantage of long-before-short relative to short-before-long orders to increase with increased length difference. This is the pattern we see in the raw mean reading times for the full sentences, which are shown in table 2. We test this prediction both for total reading time for the entire sentence, including both the target clause and the coordinated main clause that followed it, and specifically for the VP, i.e. the verb and two NPs which are the exact locus of the manipulation. The sentence reading times are reaction times (time to button press to request next sentence), while the VP reading times are total fixation duration on V, NP1 and NP2.

INSERT TABLE 2 APPROXIMATELY HERE, TABLE 3 FOLLOWING IT.

INSERT FIGURE 1 APPROXIMATELY HERE

The analysis of sentence reading time is summarised in table 3 and illustrated in figure 1; note that in the statistical model, reaction time was log transformed to reduce skewness. This showed a pattern that is compatible with the length account of Hawkins (1994), though not completely in accordance with the predictions of that account. The main result is a significant interaction between length difference and order which is driven by the fact that the difference between short-before-long (solid lines in the top left panel of figure 1) and long-before-short (dashed lines in the top left panel of figure 1) order of NPs is only significant in the case where the difference between the two NPs is more than four words ($p < 0.0001$, vs. $p = 0.2540$ for two-word

difference and 0.1674 for 4-word difference). The trend for the other two values of difference, two and four words, goes in the predicted direction, as seen in the top left panel of figure 1, but does not reach significance. This is in accordance with Hawkins' account in the sense that his account predicts an advantage for short-before-long, which we do observe, though only for the more extreme cases of length differences of more than four words. Similarly, it is in accordance with the predictions of Gibson's account which predicts a preference for short-before-long and a greater preference in the context of the extreme length difference than for the 2 or 4 word difference.

In addition to this interaction, we included a main effect of summed log conditional trigram probabilities, which was only marginally significant with $p = 0.06$. As we would expect, the effect is negative, with lower reaction times for higher probabilities. Although it is marginally significant, we include it in the analysis as an index of surprisal. The interaction between order and length difference, which shows a locality effect, was significant whether or not the trigram probability was included in the model.

The analysis also showed effects of several of the control variables, which are also illustrated in figure 1. In the top right panel, we see a non-linear effect of trial number, which reflects an initial learning or habituation effect which flattens out about halfway through the experiment. The bottom left panel shows the effect of sentence length in characters: unsurprisingly, longer sentences, counting both the target NP-construction clause and its coordinated main clause, take longer to read. This variable constitutes a formal measure of length, which turns out, in this and several of the other analyses, to show significance over and above the more syntactic conceptualisation of length that is expressed in the combination of order and difference. Finally, the bottom right panel shows the effect of repetition of sentence construction and basic referents: a clear reduction in processing time moving from the first to the second and third encounters, but not from the second to the third. In other words, the participants benefit from the reuse of the same

lexical items, but the effect is not cumulative, so when they encounter the same basic sentence (with added or removed modifiers compared to previous encounters) the third time, no extra benefit is observed. This is at first sight surprising, given that syntactic priming is typically cumulative in both production and comprehension (e.g. Branigan, 2007, Kizach and Balling 2013). It may, however, be due to the non-speeded task used here: when comprehension speed is not an issue, the gain from repeated exposure to the same basic sentence referents is not maximised in the same way as they are with a speeded task (such as the speeded acceptability task used by Kizach and Balling 2013), and the priming effect therefore is not cumulative.

INSERT TABLE 4 APPROXIMATELY HERE

A potentially more precise measure of processing difficulty is the total fixation time on the VP, i.e. the sum of all fixations on the verb, first and second NPs, the precise locus of the phenomenon of interest. The summary of the fixation times on the VP by condition are shown in table 4, and the analysis of this response variable is summarised in table 5. Here, the predicted effect of order is observed across the board, with no interaction: V-NP1-NP2 segments are always read faster when the order of NPs is short-before-long compared to long-before-short, as illustrated in the top left panel of figure 2. Similarly, the effect of length difference between the NPs was the same for both NP orders, with increasing total fixation times for segments containing longer NPs, whether they occurred in short-before-long or long-before-short order, as illustrated in the top right panel of figure 2. This is interpretable as a length effect, since the V-NP1-NP2 segments were always longer when the difference was larger; in this model, the formal measure of length in characters was not significant. In this analysis, the effect of trigram probability was solidly non-significant ($p = 0.980$) and therefore not included in the final model. Finally, this analysis showed a

non-linear effect of trial and a repetition effect, both of which were similar to that in the previous analysis, though in this analysis, the latter was only marginally significant.

INSERT TABLE 5 APPROX. HERE

INSERT FIGURE 2 AROUND HERE

Starting with the locality-based theoretical account in Hawkins (1994), we predicted two things: First, short-before-long should be faster than long-before-short, and second, the size of the length difference should increase the advantage of short-before-long order. Interestingly, the results show that the first prediction is true for the VP (short-before-long is always faster), whereas the second prediction is (somewhat) true for the total reading time. In other words, what we find is that the long-before-short order has a negative effect on processing, and this negative effect has a significant impact on the target VP for all length differences. For the entire sentence, by contrast, the size of the length difference does matter, since only a very large length difference has a significant effect on reading time.

This difference in significance may be a result of the size of the effects relative to the total times in each analysis: for length differences 2 and 4, the size of the order effect is in fact similar between the VP-analysis and the sentence analysis, namely a short-before-long advantage for length difference 2 of 78 ms for the VP and 130 ms for the sentence, and for length difference 4, 233 ms for the VP and 259 ms for the sentence; these differences are of course much larger relative to the total times for the VP-analysis than for the sentence analysis. For the biggest length difference, the advantage for short-before-long is substantially larger for the whole sentence, namely 1040 ms, compared to the VP, where it is 604 ms. Together, the two analyses suggest that there is a general

disadvantage for the non-local relations in long-before-short orders, which in the case of the extreme cases of length difference also affects reading beyond the VP itself.

With respect to the trigram probability measure, this can be seen as an index of lexicalised surprisal, as argued above. An effect of this would therefore provide evidence for a surprisal account of sentence processing; however, no such effect is observed for the VP, which is the target of the manipulation, and the effect is only marginally significant in the analysis of the reading time for the entire sentence. There are two possible interpretations of this (apart from the obvious interpretation that a null result is due to too low power to detect the effect in question): One possibility is that surprisal theory is incorrect as an overall theory of sentence processing, though this of course goes against the previous evidence in favour of this theory. Another possibility is that the problem lies with this lexicalised measure of surprisal that is the only one available to us at this level of analysis. To investigate this further, we turn to the analysis of eye movements for NP2 which is the locus for which surprisal theory makes specific predictions. The manipulation of this constitutes a syntactic conceptualisation of surprisal, in contrast to the conditional trigram probabilities which express lexicalised surprisal.

3.2 Surprisal vs. locality for processing of NP2

Surprisal theory makes a quite specific, localised prediction for our stimuli, namely that NP2 in long-before-short orders should become easier to process with increased length of NP1, i.e. with increased length difference. This prediction is based on the idea that further modification of NP1 becomes increasingly unlikely the more NP1 has already been modified, making the occurrence of the obligatory NP2 increasingly likely; this is clearly a syntactic understanding of surprisal, in contrast to the lexicalised version expressed in the trigram probability measure. To test this prediction, we analyse eye-movement data specifically for NP2 in long-before-short orders. These

NPs are always two words, while NP1 varies with length difference. We look at three different eye-movement measures: total fixation duration on NP2, first pass reading time for NP2 and regressions out of NP2. The descriptive statistics for these three measures are given in table 6. In the analysis, the latter variable is operationalised as a binary variable, because of the highly skewed distribution with a majority of zero values; it is therefore analysed using a binomial mixed model.

INSERT TABLES 6 TO 9 FROM THIS POINT.

The results of the two fixation time analyses reported in tables 7 and 8, total and first pass fixation times, are quite consistent and show no effects of difference; in other words, the reading time for NP2 is not affected by the length of NP1. This runs counter to the prediction of surprisal theory. At first sight, this absence of an effect is also in contrast with the predictions of a locality theory; however, the locality prediction is much less localised than that of the surprisal, and the effect does appear in the more global analyses reported above. The two fixation time analyses also show no effect of lexicalised surprisal as expressed in the trigram probabilities, but they do show effects of the same control variables as the more global measures, namely repetition (faster processing of occurrence 2 and 3 compared to 1, which was highly significant for sentence reading time and marginally significant for VP fixation duration) and length (longer reading times for longer NP2's).

In contrast, the binary regressions out analysis summarised in table 9 shows no effects of any of the control variables, which is not perhaps surprising given the relatively few regressions out. It does, however, show an effect of difference, with significantly lower probability of a regression out of NP2 after a long NP1 (i.e. for length difference >4). If we interpret a regression out of an NP as an attempt to integrate the NP into the syntactic structure, then one would expect to

see more regressions out, when the other NP is long, because this is the more difficult situation (when one NP is very long, the sentence as such is longer and more difficult to process, everything else being equal). But we observe the opposite pattern, suggesting that participants use less effort to integrate the NP into the structure in precisely the situation where this is most difficult. This finding could be taken as support for the so-called good-enough approach to parsing (Ferreira, Bailey, & Ferraro, 2002; Ferreira & Patson, 2007), where the central idea is that processing becomes more shallow when difficulties are encountered. The idea is that rapid interpretation of input is more important than deep understanding of input, so when the processing speed is threatened by complex input, the parser will adjust the depth to maintain the speed, but at the cost of precision. In the vicinity of a long NP, it may be the case that less attention is given to the short and easy NP to maintain a reasonable processing speed. This in accordance with previous results from Danish showing that information structural preferences (new-before-given) are neutralized in a long-before-short context (Kizach & Balling, 2013), suggesting precisely that the depth of understanding is reduced under more difficult circumstances.

Another way of interpreting the fewer occurrences of regressions out of NP2 when NP1 is very long is as a sign that processing is simply easier in these cases. If the occurrence of NP2 is highly predictable following an extremely long NP1, then surprisal, in the syntactic definition we employ in the length difference variable, would precisely predict that processing is easier. However, it would be odd (or even surprising) if surprisal only manifests as a reduction in regressions out and has no reflex in the other measures used.

4. Conclusion

Summing up, we see strong and consistent effects of locality, with only minor deviations from the predictions, while for surprisal it is rather the other way around: we find only marginal effects that

are consistent with a surprisal account, while the deviations from the predictions are major. Turning first to the pattern of locality effects, we see that across both reading of the entire sentence (the target clause and its coordinated main clause) and specifically for the VP, there is a disadvantage for the long-before-short orders relative to short-before-long. For the entire sentence, this is significant only for the largest length difference, with an effect size of approximately one second, while for the VP, the order effect is statistically speaking the same across the three length differences. This pattern of results confirm the key predictions arising from the locality accounts of Hawkins (1994, 2004, 2014) and Gibson (1998, 2003), that long-before-short orders should be at a disadvantage because the syntactic relations are less local.

As for surprisal, there were two results that could be interpreted in favour of a surprisal account and both of these were marginal. The first was the effect of conditional trigram probability, indexing lexicalised surprisal, which was marginally significant. We found this in the analysis of the most global variable, reading time for the entire sentence, which is contrary to what we would expect, since surprisal theory makes the most specific predictions not at the global level of the entire sentence but at the precise locus of the onset of NP2. The other possible effect of surprisal is the effect of the length difference on regressions out of NP2 in long-before-short orders: the fewer regressions out for the biggest length difference could perhaps be interpreted in terms of easier processing of that NP2 when the length difference is big, but it remains peculiar that the effect is only seen for the most ambiguous measure of difficulty, namely regressions out. In contrast to the other possible surprisal effect, this is syntactic rather than lexicalised surprisal.

In other words, neither the syntactic nor the lexicalised index of surprisal showed any strong or consistent effects. This largely null effect of surprisal may of course be a question of statistical power, but the fact that we do find consistent effects of several other variables points in the direction that the surprisal effect is, if not absent for Danish and/or for this type of construction,

then at least not very strong. Consistently with this, previous evidence of anti-locality and surprisal comes mainly from analyses of large eye-movement corpora (the Dundee corpus in the case of Demberg and Keller (2008) and Frank and Bod (2011); the Potsdam corpus in the case of Boston et al. (2008)) and from experiments with verbs in verb-final languages (Konieczny, 2000; Konieczny & Döring, 2003; Levy & Keller, 2013; Smith & Levy, 2013; Vasishth & Lewis, 2006) . Verbs in verb-final languages may be a particularly strong case because of the central role of the verb in the sentence structure and the expectations that therefore may be established before encountering a verb in a verb-final sentence. In short, it may be the case that surprisal effects are absent or minor in most cases, but relatively strong for verbs in verb-final languages. Another possibility is of course that surprisal effects are not relevant for Danish, but given the effects found for the genetically and typologically closely related languages English and German, this seems an unlikely explanation. In short, the main surprise in this experiment is the lack of surprisal effects – locality on the other hand gives us no surprises: as expected, local syntactic relations are systematically preferred.

5. References

Balling, L. W. (2013). Reading authentic texts: What counts as cognate? *Bilingualism: Language and Cognition*, *16*(3), 637–653.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using Eigen and S4 (Version 1.1-10). Retrieved from http://cran.r-project.org/web/packages/lme4/index.html

Boston, M., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *The Mind Research Repository (Beta)*, (1). Retrieved from http://openscience.uni-leipzig.de/index.php/mr2/article/view/62

Branigan, H. (2007). Syntactic priming. *Language and Linguistics Compass*, *1*(1–2), 1–16.

Brown, M., Savova, V., & Gibson, E. (2012). Syntax encodes information structure: Evidence from on-line reading comprehension. *Journal of Memory and Language*, *66*(1), 194–209. https://doi.org/10.1016/j.jml.2011.08.006

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359–394.

Christensen, K. R., Kizach, J., & Nyvad, A. M. (2013). Escape from the Island: Grammaticality and (Reduced) Acceptability of wh-island Violations in Danish. *Journal of Psycholinguistic Research*, *42*(1), 51–70.

Clifton, C., & Frazier, L. (2004). Should given information come before new? Yes and no. *Memory & Cognition*, *32*(6), 886–895.

De Cuypere, L., & Verbeke, S. (2013). Dative alternation in Indian English: A corpus-based analysis. *World Englishes*, *32*(2), 169–184.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15.

Ferreira, F., & Patson, N. D. (2007). The ?Good Enough? Approach to Language Comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Fodor, J. D., & Inoue, A. (1998). Attach anyway. In J. D. Fodor & F. Ferreira (Eds.), *Reanalysis in sentence processing* (pp. 101–141). Springer.

Fodor, J. D., & Inoue, A. (2000). Garden path re-analysis: Attach (anyway) and revision as last resort. In V. Lombardo (Ed.), *Cross-linguistic perspectives on language processing* (pp. 21–61). Springer.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). CA: Sage.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834.

Frazier, L. (1987). Sentence Processing: a Tutorial Review. In M. Coltheart (Ed.), *Attention and Performance XII*. Hove and London / Hillsdale: Lawrence Erlbaum Associates.

Frazier, L., & Clifton, C. (1996). *Construal*. MIT Press, Cambridge.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium* (pp. 95–126). Cambridge, MA: MIT Press.

Gibson, E. (2003). Sentence Comprehension, Linguistic Complexity in. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science*. UK: Nature Publishing Group.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1–8). Association for Computational Linguistics.

Hawkins, J. A. (1990). A Parsing Theory of Word Order Universals. *Linguistic Inquiry*, *21*(2), 223–261.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, J. A. (1998). A processing approach to word order in Danish. *Acta Linguistica Hafniensia*, *30*(1), 63–101.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hawkins, J. A. (2011). Discontinuous dependencies in corpus selections: Particle verbs and their relevance for current issues in language processing. In E. M. Bender & J. E. Arnold (Eds.), *Language from a cognitive perspective: grammar, usage and processing* (pp. 269–290). Cambridge, MA: CLSI Publications.

Hawkins, J. A. (2014). *Cross-linguistic Variation and Efficiency*. Oxford: Oxford University Press.

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, *86*(2), 366–415. https://doi.org/10.1353/lan.0.0223

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing* (2nd Edition). Upper Saddle River, New Jersey: Pearson Education Inc.

Kizach, J. (2012). Evidence for weight effects in Russian. *Russian Linguistics*, *36*(3), 251–270. https://doi.org/10.1007/s11185-012-9096-0

Kizach, J. (2014). A multifactorial analysis of the Russian adversity impersonal construction. *Russian Linguistics*, *38*(2), 205–211. https://doi.org/10.1007/s11185-014-9128-z

Kizach, J. (2015). Animacy and the ordering of postverbal prepositional phrases in Danish. *Acta Linguistica Hafniensia*, 1–21.

Kizach, J., & Balling, L. W. (2013). Givenness, complexity, and the Danish dative alternation. *Memory & Cognition*, *41*(8), 1159–1171. https://doi.org/10.3758/s13421-013-0336-3

Kizach, J., & Vikner, S. (2016). Head adjacency and the Danish dative alternation. *Studia Linguistica*. https://doi.org/10.1111/stul.12047

Koehn, P. (2010). *Statistical machine translation*. Cambridge: Cambridge University Press.

Konieczny, L. (2000). Locality and Parsing Complexity. *Journal of Psycholinguistic Research*, *29*(6), 627–645. https://doi.org/10.1023/A:1026528912821

Konieczny, L., & Döring, P. (2003). Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. In *Proceedings of iccs/ascs*.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package "lmerTest" (Version 2.0-29). Retrieved from http://CRAN.R-project.org/package=lmerTest

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R., Fedorenko, E., & Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *Journal of Memory and Language*, *69*(4), 461–495.

Levy, R., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*, *68*(2), 199–222.

MacDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research*, *43*, 1735–1751.

Pritchett, B. L. (1992). *Grammatical competence and parsing performance*. University of Chicago Press.

R Development Core Team. (2014). R: A language and environment for statistical computing (Version 3.1.1). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rosenbach, A. (2005). Animacy Versus Weight as Determinants of Grammatical Variation in English. *Language*, *81*(3), 613–644. https://doi.org/10.1353/lan.2005.0149

Seoane, E. (2009). Syntactic complexity, discourse status and animacy as determinants of grammatical variation in Modern English. *English Language and Linguistics*, *13*(3), 365. https://doi.org/10.1017/S1360674309990153

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Stolcke, A., Zheng, J., Wang, W., & Abrash, V. (2011). SRILM at sixteen: Update and outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop* (Vol. 5).

Szmrecsanyi, B. (2004). On operationalizing syntactic complexity. *Jadt-04*, *2*, 1032–1039.

Thornton, R., MacDonald, M. C., & Arnold, J. E. (2000). The concomitant effects of phrase length and informational content in sentence comprehension. *Journal of Psycholinguistic Research*, *29*(2), 195–203.

Vasishth, S., & Lewis, R. L. (2006). Argument-Head Distance and Processing Complexity: Explaining both Locality and Antilocality Effects. *Language*, *82*(4), 767–794. https://doi.org/10.1353/lan.2006.0236

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, *85*(1), 79–112.

Wasow, T. (1997). Remarks on grammatical weight. *Language Variation and Change*, *9*(1), 81–105.

Wasow, T. (2002). *Postverbal Behavior*. Stanford: CSLI Publications.

Wiechmann, D., & Lohmann, A. (2013). Domain minimization and beyond: Modeling prepositional phrase ordering. *Language Variation and Change*, *25*(1), 65–88.

Table 1: A sample set of stimulus sentences in the experiment

Difference	Order	
	Short-before-long	Long-before-short
2 words	[Butiksindehaveren]$_{S}$ [gav]$_{V}$ [en brudepige]$_{NP1/IO}$ [en kjole fra Italien]$_{NP2/DO}$	[Butiksindehaveren]$_{S}$ [gav]$_{V}$ [en brudepige fra Italien]$_{NP1/IO}$ [en kjole]$_{NP2/DO}$
	[The shopkeeper]$_{S}$ [gave]$_{V}$ [a bridesmaid]$_{NP1/IO}$ [a dress from Italy]$_{NP2/DO}$	[The shopkeeper]$_{S}$ [gave]$_{V}$ [a bridesmaid from Italy]$_{NP1/IO}$ [a dress]$_{NP2/DO}$
4 words	[Butiksindehaveren]$_{S}$ [gav]$_{V}$ [en brudepige]$_{NP1/IO}$ [en meget flot kjole fra Italien]$_{NP2/DO}$	[Butiksindehaveren]$_{S}$ [gav]$_{V}$ [en meget flot brudepige fra Italien]$_{NP1/IO}$ [en kjole]$_{NP2/DO}$
	[The shopkeeper]$_{S}$ [gave]$_{V}$ [a bridesmaid]$_{NP1/IO}$ [a very good-looking dress from Italy]$_{NP2/DO}$	[The shopkeeper]$_{S}$ [gave]$_{V}$ [a very good-looking bridesmaid from Italy]$_{NP1/IO}$ [a dress]$_{NP2/DO}$
More words	[Butiksindehaveren]$_{S}$ [gav]$_{V}$ [en brudepige]$_{NP1/IO}$ [en kjole fra Italien der lige var ankommet]$_{NP2/DO}$	[Butiksindehaveren]$_{S}$ [gav]$_{V}$ [en meget flot brudepige fra Italien der lige var ankommet,]$_{NP1/IO}$ [en kjole]$_{NP2/DO}$
	[The shopkeeper]$_{S}$ [gave]$_{V}$ [a bridesmaid]$_{NP1/IO}$ [a very good-looking dress from Italy that had just arrived]$_{NP2/DO}$	[The shopkeeper]$_{S}$ [gave]$_{V}$ [a very good-looking bridesmaid from Italy that had just arrived]$_{NP1/IO}$ [a dress]$_{NP2/DO}$

A basic sentence with variations in ordering (short before long vs. long before short) and difference (two-word, four-word or more words length difference between NP1 and NP2).

Table 2: Mean sentence reading times in ms by condition (sd's in parentheses)

	Short-before-long	Long-before-short
Difference 2 words	4756 (1744)	4887 (1800)
Difference 4 words	5112 (1550)	5371 (1982)
Difference more words	5910 (1920)	6950 (2667)

Table 3: Analysis of sentence reading time

Fixed effects

	Estimate	Std. error	df	t	p
Intercept	8.0250	0.0961	106.8	83.5060	< 0.0001
Trial (linear, scaled)	-0.1120	0.0258	1241.0	-4.3500	< 0.0001
Trial (quadratic, scaled)	0.0879	0.0252	1245.0	3.4850	0.0005
Sentence length in characters	0.0026	0.0008	81.1	3.0580	0.0030
Repetition: 2	-0.1006	0.0153	1234.0	-6.5770	< 0.0001
Repetition: 3	-0.1515	0.0190	1227.0	-7.9810	< 0.0001
Sum log trigram probs for sentence	-0.0034	0.0018	80.8	-1.8870	0.0628
Difference: 4words	0.0408	0.0294	81.3	1.3890	0.1687
Difference: morewords	0.1029	0.0388	83.9	2.6490	0.0096
Difference2words:long-before-short	0.0326	0.0284	83.7	1.1490	0.2540
Difference4words:long-before-short	0.0395	0.0284	84.0	1.3930	0.1674
Differencemorewords:long-before-short	0.1401	0.0284	84.5	4.9290	< 0.0001

Random effects

Groups	Name	Variance	Std.Dev.	Correlations		
ID	Intercept	0.003117	0.05583			
Participant	Intercept	0.051183	0.22624			
	Order:long-before-short	0.001783	0.04222	0.42		
	Difference: 4words	0.000503	0.02242	-0.98	-0.59	
	Difference: morewords	0.001743	0.04174	-0.18	0.82	-0.02
Residual		0.035465	0.18832			

Summary of linear mixed-effects model of reading times on target sentences. Factors are treatment coded with the reference levels 1 for the factor Repetition and 2 words for the factor Difference. For the factor Order, we parametrized the model to get tests of the Order effect separately for the three levels of Difference. The model is one where 27 observations (2%) with large standardised residuals (above an absolute value of 2.5) were removed.

Table 4: Mean fixation time in ms on the VP, by condition (sd's in parentheses)

	Short-before-long	Long-before-short
Difference 2 words	2001 (1001)	2079 (1033)
Difference 4 words	2344 (927)	2577 (1276)
Difference more words	3118 (1300)	3722 (1797)

Table 5: Analysis of fixations on VP

Fixed effects

	Estimate	Std. error	df	t	p
Intercept	7.4933	0.0655	37.9	114.4270	< 0.0001
Trial (linear, scaled)	-0.1002	0.0377	1209.9	-2.6560	0.0080
Trial (quadratic, scaled)	0.0762	0.0370	1215.8	2.0600	0.0396
Repetition: 2	-0.0473	0.0259	57.2	-1.8290	0.0725
Repetition: 3	-0.0648	0.0333	70.5	-1.9480	0.0554
Order: Long-short	0.0889	0.0287	76.2	3.1050	0.0027
Difference: 4words	0.2174	0.0356	77.8	6.1160	0.0000
Difference: morewords	0.5522	0.0380	66	14.5430	< 0.0001

Random effects

Groups	Name	Variance	Std.Dev.	Correlations				
ID	Intercept	0.0108	0.1037					
Participant	Intercept	0.1008	0.3174					
	Repetition: 2	0.0053	0.0727	0.23				
	Repetition: 3	0.0104	0.1019	-0.02	0.68			
	Order: Long-short	0.0033	0.0578	0.2	-0.47	0.13		
	Difference: 4words	0.0059	0.0771	-0.87	-0.15	0.13	0.07	
	Difference: morewords	0.0112	0.1060	-0.59	-0.21	-0.05	0.24	0.89
Residual		0.0741	0.2723					

Summary of linear mixed-effects model of total fixation duration on target region (verb and NPs). Factors are treatment coded with the reference levels 1 for the factor Repetition, Short-before-long for Order and 2 words for the factor Difference. The model is one where 21 observations (1.6%) with large standardised residuals (above an absolute value of 2.5) were removed.

Table 6: Mean total and first pass fixation times and regressions out of NP2, by condition (sd's in parentheses)

	Total fixation time, ms	First pass fixation time, ms	Regressions out, count
Difference 2 words	552 (351)	386 (230)	0.49 (0.70)
Difference 4 words	519 (336)	384 (225)	0.42 (0.74)
Difference more words	522 (309)	397 (190)	0.29 (0.55)

Table 7: Analysis of fixations on NP2

Fixed effects

	Estimate	Std. error	df	t	p
Intercept	5.7846	0.1668	51.47	34.6720	<0.0001
NP2 length in characters	0.0453	0.0146	38.96	3.0910	0.0037
Repetition: 2	-0.1469	0.0484	138.68	-3.0350	0.0029
Repetition: 3	-0.2118	0.0575	29.89	-3.6830	0.0009
Difference: 4words	-0.0349	0.0777	40.74	-0.4490	0.6559
Difference: morewords	-0.0094	0.0805	39.64	-0.1170	0.9075

Random effects

Groups	Name	Variance	Std.Dev.	Correlations			
ID	Intercept	0.0276	0.1660				
Participant	Intercept	0.1055	0.3248				
	Repetition: 2	0.0053	0.0727	0.99			
	Repetition: 3	0.0356	0.1886	0.34	0.43		
	Difference: 4words	0.0092	0.0959	-0.98	-0.97	-0.22	
	Difference: morewords	0.0216	0.1470	-0.73	-0.69	-0.38	0.59
Residual		0.2139	0.4625				

Summary of linear mixed-effects model of total fixation duration on NP2. Factors are treatment coded with the reference levels 1 for the factor Repetition and 2 words for the factor Difference.

Table 8: Analysis of first pass reading time for NP2

Fixed effects

	Estimate	Std. error	df	t	p
Intercept	4.9978	0.2492	29.7	20.0530	<0.0001
NP2 length in characters	0.0491	0.0093	37.2	5.2810	<0.0001
Repetition: 2	-0.1302	0.0411	590.8	-3.1640	0.0016
Repetition: 3	-0.1753	0.0409	587.2	-4.2840	<0.0001
Sum log trigram probs for NP2	-0.0676	0.0379	28.7	-1.7840	0.0849
Difference: 4words	-0.0224	0.0504	32.6	-0.4440	0.6601
Difference: morewords	0.0623	0.0502	26.1	1.2410	0.2256

Random effects

Groups	Name	Variance	Std.Dev.	Correlations		
ID	Intercept	0.004398	0.06632			
Participant	Intercept	0.066703	0.25827			
	Sum log trigram probs for NP2	0.001164	0.03412	0.54		
	Difference: 4words	0.009618	0.09807	-0.71	0.2	
	Difference: morewords	0.008932	0.09451	0.48	0.82	0.17
Residual		0.170474	0.41289			

Summary of linear mixed-effects model of first pass reading time on individual NPs. Factors are treatment coded with the reference levels 1 for the factor Repetition and 2 words for the factor Difference. In this model, 12 observations (1.9%) with large standardised residuals (above an absolute value of 2.5) were removed.

Table 9: Analysis of regressions out

Fixed effects

	Estimate	Std. Error	z	p
Intercept	-0.6327	0.2779	-2.277	0.0228
Difference4	-0.3042	0.3315	-0.918	0.3588
Differencemange	-0.7934	0.3384	-2.344	0.0191

Random effects

Groups	Name	Variance	Std.Dev.
ID	Intercept	0.4458	0.6677
Participant	Intercept	0.6666	0.8165

Summary of logistic mixed-effects model of regressions from individual NPs (formulated as a binary variable: +/- regressions out). The factor Difference is treatment coded with the reference level 2 words.