Data from Eye-tracking Corpora as Evidence for Theories of Syntactic Processing Complexity

Vera Demberg and Frank Keller
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, UK
phone: +44-131-650-4407, fax: +44-131-650-6626
email: v.demberg@ed.ac.uk, frank.keller@ed.ac.uk

Abstract

We evaluate the predictions of two theories of syntactic processing complexity, dependency locality theory (DLT) and surprisal, against the Dundee corpus, which contains the eye-tracking record of 10 participants reading 51,000 words of newspaper text. Our results show that DLT integration cost is not a significant predictor of reading times for arbitrary words in the corpus. However, DLT successfully predicts reading times for nouns and verbs. We also find evidence for integration cost effects at auxiliaries, not predicted by DLT. For surprisal, we demonstrate that an unlexicalized formulation of surprisal can predict reading times for arbitrary words in the corpus. Comparing DLT integration cost and surprisal, we find that the two measures are uncorrelated, which suggests that a complete theory will need to incorporate both aspects of processing complexity. We conclude that eye-tracking corpora, which provide reading time data for naturally occurring, contextualized sentences, can complement experimental evidence as a basis for theories of processing complexity.

Keywords: eye-tracking, corpus data, processing complexity, dependency locality theory, surprisal

---

This research was supported by EPSRC grant EP/C546830/1 “Prediction in Human Parsing”. We are grateful to Roger Levy for important suggestions and comments regarding this research. Fernanda Ferreira, Patrick Sturt, Manabu Arai, and two anonymous reviewers have provided valuable feedback on an earlier version of this paper. Asaf Bachrach provided a text that was manually annotated with integration cost scores, which enabled us to evaluate our implementation. Brian Roark kindly made his incremental parser available for this work, and even modified it to enable the computation of prefix probablities.

1. Introduction

Research on human sentence processing has traditionally focused on syntactic ambiguity, based on the observation that certain locally ambiguous constructions pose difficulty for the human sentence processor. Such difficulty manifests itself typically in the form of increased processing time (e.g., elevated reading times on the disambiguating region).

While disambiguation is an important source of difficulty in human sentence processing, difficulty can also arise in unambiguous sentences. A classic example are relative clauses, which have been investigated extensively in the literature on syntactic processing difficulty. Experimental results show that English subject relative clauses as in (1-a) are easier to process than object relative clauses as in (1-b). Experimentally, this difficulty is evidenced by the fact that reading times for the verb *attacked* are shorter for subject relative clauses than for object relative clauses (King & Just, 1991).

(1) a. The reporter who attacked the senator admitted the error.
b. The reporter who the senator attacked admitted the error.

Findings such as these have motivated processing theories that do not rely on ambiguity resolution, but instead capture the complexity involved in computing the syntactic dependencies between the words in a sentence. One such theory is Dependency Locality Theory (DLT), proposed by Gibson (1998, 2000). A central notion in DLT is *integration cost*, a distance-based measure of the amount of processing effort required when the head of a phrase is integrated with its syntactic dependents. DLT is able to capture the subject/object relative clause asymmetry in (1), as well as a wide range of other complexity results, including processing overload phenomena such as center embedding and cross-serial dependencies.

More recently, Hale (2001) proposed surprisal as an alternative measure of processing complexity. Intuitively, the surprisal of a word in a sentential context corresponds to the probability mass of the analyses that are not consistent with the new word. Surprisal requires a probabilistic notion of linguistic structure (utilizing transitional probabilities or probabilistic grammars), and has its theoretical foundation in information theory (Levy, 2008). It can be shown to capture a range of complexity effects, including the subject/object relative clause asymmetry, certain garden path effects, speed-up effects in verb-final contexts, and word order asymmetries in German (Hale, 2001; Levy, 2008). Another more recent incarnation of surprisal Gibson's (2006) approach, which combines top-down syntactic predictions with bottom-up lexical predictions.

A number of other theories of syntactic processing complexity exist, including dynamic system models (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998; Tabor & Tanenhaus, 1999) and neural net models (e.g., Elman, 1991). However, in the present paper, we will focus on DLT and surprisal, as these two approaches are maximally different from each other. In particular, they make complementary assumptions about the source of processing complexity. DLT's integration cost captures the cost incurred when a head has to be integrated with the dependents that precede it. Surprisal, on the other hand, accounts for the cost that results when the current word is not predicted by the preceding context. Therefore, integration cost can be regarded as a backward looking cost (past material has to be held in memory and integrated), while surprisal is a forward-looking cost (syntactic predictions have to be discarded if they are no longer compatible with the current word). This observation leads to a general empirical prediction, viz., that integration cost and surprisal should be uncorrelated, and should account for complementary aspects of reading time data. The present

paper will test this prediction.

While DLT and surprisal have been evaluated against a range of experimental results, so far no *broad coverage* evaluation of theories of syntactic processing complexity has been carried out. Existing studies rely on lab experiments, which have the advantage of giving the experimenter full control over the experimental setup and the materials, and are of established reliability and validity. However, this approach also has its drawbacks. It typically involves the presentation of artificially constructed sentences containing a narrow range of syntactic structures. Also, the same structures occur many times in a given experiment, raising the possibility of habituation effects or the development of strategies in participants. The sentences to be tested are often presented in isolation, i.e., without an appropriate textual context, potentially leading to behavior that is different from normal reading. Finally, only a small number of items can be tested in the typical psycholinguistic experiment. DLT and surprisal effects have successfully obtained in such an experimental setting, but these methodological limitations leave open the possibility that the effects are rare or absent when arbitrary words in large amounts of naturalistic, contextualized text are considered.

The aim of the present paper is to address this problem and provide a broad coverage evaluation of DLT and surprisal on the Dundee Corpus, a large collection of newspaper text for which the eye-movement record of 10 participants is available. From this corpus, a range of eye-tracking measures can be computed, which can then be evaluated against the predictions of theories of syntactic complexity. Such broad coverage studies yield results that hold for naturalistic, contextualized text, rather than for isolated example sentences artificially constructed by psycholinguists. They have already been applied successfully to individual phenomena, such as the subject/object relative clause asymmetry (Demberg & Keller, 2007). The aim of the present paper is to show that corpus studies can also be used to systematically test theories of syntactic processing complexity. Such studies provide a source of evidence that corroborates experimental results, but also yields new theoretical insights, as it makes it possible to evaluate multiple theoretical predictors against each other on a large, standardized data set.

2. Background

2.1. Dependency Locality Theory

According to Gibson's (1998, 2000) Dependency Locality Theory, processing complexity is caused by the cost of the computational resources consumed by the processor. Two distinct cost components can be distinguished: (i) *integration cost* associated with integrating new input into the structures already built at a given stage in the computation, and (ii) *memory cost* involved in the storage of parts of the input that may be used in parsing later parts of an input. Here, we will focus on integration cost, as "reasonable first approximations of comprehension times can be obtained from the integrations costs alone, as long as the linguistic memory storage used is not excessive at these integration points" (Gibson, 1998, p. 19f). This is a safe assumption for our studies, as we use corpora of carefully edited newspaper text, which are unlikely to incur excessive storage costs (in contrast to artificially constructed experimental materials). Gibson's definition of integration is as follows:

(2) **Linguistic Integration Cost**
The integration cost associated with integrating a new input head h_2 with a head h_1 that is part of the current structure for the input consists of two parts: (1) a cost dependent on

> the complexity of the integration (e.g. constructing a new discourse referent); plus (2) a distance-based cost: a monotone increasing function I(n) energy units (EUs) of the number of new discourse referents that have been processed since h_1 was last highly activated. For simplicity, it is assumed that I(n) = n EUs. (Gibson, 1998, p. 12f)

According to this definition, integration cost is dependent on two factors. First, the type of element to be integrated matters: new discourse referents (e.g., indefinite NPs) are assumed to involve a higher integration cost than old/established discourse referents, identified by pronouns. Second, integration cost is sensitive to the distance between the head being integrated and the head it attaches to, where distance is calculated in terms of intervening discourse referents.

As an example, consider the subject vs. object relative clause example in (1). At the embedded verb *attacked* in (1-a), two integrations take place: the gap generated by the relative pronoun *who* needs to be integrated with the verb. The cost for this is I(0), as zero new discourse referents have been processed since the gap was encountered. In addition, the embedded verb *attacked* needs to be integrated with its preceding subject. Again, this is a free integration since no discourse referent occurs between the verb and the subject NP. However, there is a cost for building a new discourse referent (the embedded verb itself[1]), leading to a cost of I(1). The total cost at *attacked* is therefore I(1). This is illustrated in Figure 1, which depicts the dependencies that are built, and the integration costs per word that are incurred.

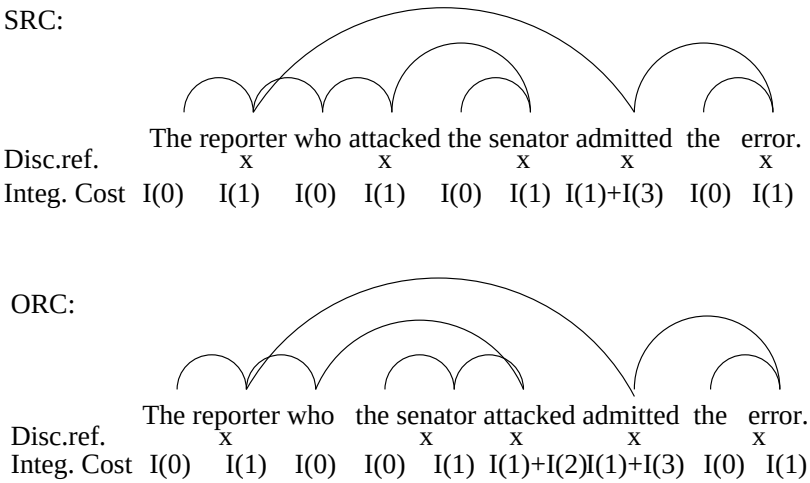


Figure 1. An example of integration cost computations: subject relative clauses (SRC) vs. object relative clauses (ORC), with word-by-word markup for discourse referent and integration costs. The links between the words represent syntactic dependencies.

At the verb *attacked* in the object relative clause, three structural integrations take place: (1) integration with the subject NP *the senator*: no integration costs occur since no new discourse referents occur in between the verb and the NP, (2) an empty category for the relative pronoun is integrated, but again, the integration is local and no costs occur, (3) the object position empty category is co-indexed with the preceding relative pronoun *who*. There is an integration cost of I(2) for this step due to the two discourse referents, *attacked* and *the senator* which occurs in between.

[1]DLT assumes that verb introduce event discourse referents.

In addition, there is a cost of I(1) for constructing the discourse referent at *attacked*, which leads to a total integration cost of I(1) + I(2) at the embedded word of the object relative clause. So overall, DLT predicts that the verb of object relative clauses is more difficult to process than that of subject relative clauses.

Note that Gibson assumes that the integration cost function is identity, i.e., I(n) = n. However, other functions are possible here; we will return to this issue in Section 3.2.

2.2. Surprisal

An alternative measure of syntactic complexity has been proposed by Hale (2001) in the form of surprisal. Surprisal is compatible with a parallel parser, which builds structures incrementally, i.e., it constructs all possible syntactic analyses compatible with the input string on a word-by-word basis.[2] Intuitively, surprisal measures the change in probability mass as structural predictions are disconfirmed when a new word is processed. If the new word disconfirms predictions with a large probability mass (high surprisal), then high processing complexity is predicted, corresponding to increased reading time. If the new word only disconfirms predictions with a small probability mass (low surprisal), then we expect low processing complexity and reduced reading time.

Returning to (1), we expect differences in surprisal between (1-b) and (1-a). Hale (2001) demonstrates that the mean surprisal for the object relative clause is higher than for the subject relative clause, i.e., that on average, the words in the object relative clause require hypotheses with a greater probability mass to be disconfirmed than in the subject relative clause. Surprisal theory therefore predicts that object relative clauses are harder to process than subject relative clauses, which is in line with experimental findings (but see Levy, 2008, for additional relative clause results using surprisal).

Technically, surprisal can be defined using the conditional probability $P(T|w_1 \cdots w_k)$, i.e., the probability of a tree T given the sentence prefix $w_1 \cdots w_k$. This is the probability that T is the correct tree, given that the string of word w_1 to word w_k has been encountered. Surprisal is then defined as the change in the conditional probability distribution from w_k to w_{k+1}. As Levy (2008) shows, this can be formalized using the Kullback-Leibler divergence (relative entropy). The Kullback-Leibler divergence between two probability distributions P and Q is defined as:

$$(1) \qquad D(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

The surprisal at encountering word w_{k+1} then corresponds to the Kullback-Leibler divergence between $P(T|w_1 \cdots w_{k+1})$, i.e., the probability distribution of all syntactic trees that are consistent with words $w_1 \cdots w_{k+1}$, and $P(T|w_1 \cdots w_k)$, the probability distribution of the trees that are compatible with the prefix $w_1 \cdots w_k$:

$$(2) \qquad S_{k+1} = \sum_T P(T|w_1 \cdots w_{k+1}) \log \frac{P(T|w_1 \cdots w_{k+1})}{P(T|w_1 \cdots w_k)}$$

This expression can be simplified using the following fact:

$$(3) \qquad P(T|w_1 \cdots w_k) = \frac{P(T, w_1 \cdots w_k)}{P(w_1 \cdots w_k)} = \frac{P(T)}{P(w_1 \cdots w_k)}$$

[2] While surprisal is compatible with a fully parallel parser, it does not necessarily require one. It is possible to compute the probabilities of a limited set of analyses and then use these to track changes in the probability distribution. In fact, the Roark (2001) parser used in this paper performs beam-search, i.e., does not compute all possible analyses, and thus we reply on such a limited-parallelism version of surprisal.

This equation holds because we know that each tree in T contains the words $w_1 \cdots w_k$, therefore $P(T, w_1 \cdots w_k) = P(T)$. We can now substitute Equation (3) into Equation (2). We can then simplify the definition of surprisal using the fact $\sum_T \frac{P(T)}{P(w_1 \cdots w_{k+1})} = 1$ (the probabilities of all syntactic trees given a particular prefix sum up to one), and performing some straightforward logarithmic transformations:

$$
\begin{aligned}
(4) \qquad S_{k+1} &= \sum_T \frac{P(T)}{P(w_1 \cdots w_{k+1})} \cdot \log \frac{\frac{P(T)}{P(w_1 \cdots w_{k+1})}}{\frac{P(T)}{P(w_1 \cdots w_k)}} = 1 \cdot \log \frac{P(w_1 \cdots w_k)}{P(w_1 \cdots w_{k+1})} \\
&= -\log \frac{P(w_1 \cdots w_{k+1})}{P(w_1 \cdots w_k)} = -\log P(w_{k+1} | w_1 \cdots w_k)
\end{aligned}
$$

This derivation shows that the surprisal S_{k+1} at word w_{k+1} corresponds to the negative logarithm of the conditional probability of w_{k+1} given the sentential context $w_1 \cdots w_k$. This is an important simplification, as it means that surprisal can be computed without making representational assumptions (i.e., the syntactic tree T does not figure in the definition of surprisal). In practice this means that a number of ways of computing surprisal are possible, utilizing either simple probabilistic models of language (such as *n*-gram models) or more sophisticated ones, such as probabilistic context-free grammars (PCFGs).

Surprisal can be reformulated in terms of the *prefix probabilities* of words w_k and w_{k+1}, which can be obtained easily from a PCFG. The prefix probability of a word w_k is obtained by summing the probabilities of all trees T that span from w_1 to w_k:

$$
(5) \qquad P(w_1 \cdots w_k) = \sum_T P(T, w_1 \cdots w_k)
$$

The formulation in Equation (4) is therefore equivalent to a formulation that uses prefix probabilities:

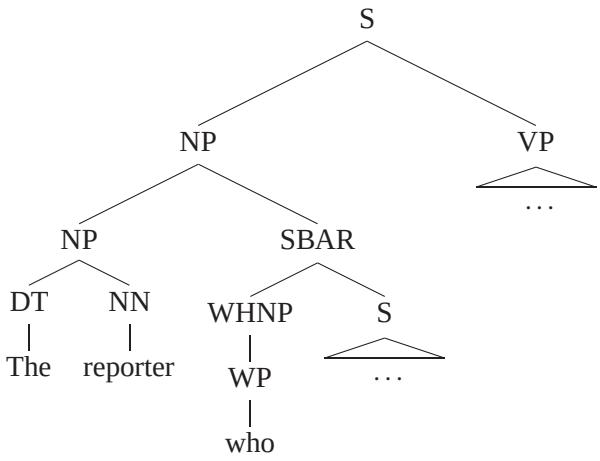
$$
(6) \qquad S_{k+1} = -\log \frac{P(w_1 \cdots w_{k+1})}{P(w_1 \cdots w_k)} = \log \sum_T P(T, w_1 \cdots w_k) - \log \sum_T P(T, w_1 \cdots w_{k+1})
$$

Surprisal S_{k+1} at word w_{k+1} thus corresponds to the difference between the logarithm of the prefix probabilities of word w_k and w_{k+1}. We give an example that illustrates how prefix probabilities can be computed using a PCFG. In a PCFG, each context-free grammar rule is annotated with its probability, as in Figure 2. The rule probabilities are then used to calculate the prefix probability of a word.

For example, if w_{k+1} is the word *who* in the example in Figure 2, then the prefix probability $\sum_T P(T, w_1 \cdots w_{k+1})$ is the sum over the probabilities of all possible trees that include the prefix $w_1 \cdots w_{k+1}$, where each tree probability is computed as the product of all the rules that are needed to build the tree (Figure 2 shows only one such tree).

2.3. Non-syntactic Predictors

It is well-known that reading times in eye-tracking data are influenced not only by high-level, syntactic variables but also by a number of low-level variables, both linguistic ones and oculomotor ones (see Rayner, 1998, for a review). The low-level linguistic variables include word frequency (more frequent words are read faster), word length (shorter words are read faster), and the position of the word in the sentence (later words are read faster). It has also been found that the frequency



Example	Rule	Rule probability
The reporter who ...	S → VP NP	$p = 0.6$
The reporter who ...	NP → NP SBAR	$p = 0.004$
The reporter	NP → DT NN	$p = 0.5$
The	DT → the	$p = 0.7$
reporter	NN → reporter	$p = 0.0002$
who ...	SBAR → WHNP S	$p = 0.12$
who	WHNP → WP	$p = 0.2$
who	WP → who	$p = 0.8$

Figure 2. Example derivation of prefix *The reporter who* and rules from a probabilistic context free grammar (PCFG) that would be needed in order to calculate its prefix probability.

of the previous word influences reading time at the present word, presumably due to parafoveal preview. Oculomotor variables include previous fixation (indicating whether or not the previous word has been fixated), launch distance (how many character intervene between the current fixation and the previous fixation), and landing position (which letter in the word the fixation landed on).

Together with variation between readers, these low-level variables account for a sizable proportion of the variance in the eye-movement record. There are also a number of well-known correlations between the independent variables: short words are usually more frequent than long words, the fixation landing position depends on word length, etc.

Recently, it has also been shown that information about the sequential context of a word can influence reading times. In particular, McDonald and Shillcock (2003b) present data extracted from an eye-tracking corpus (a smaller corpus than the Dundee Corpus used here) that show that forward and backward transitional probabilities are predictive of first fixation and first pass durations: the higher the transitional probability, the shorter the fixation time.

By *forward transitional probability* McDonald and Shillcock (2003b) refer to the conditional probability of a word given the previous word $P(w_k|w_{k-1})$. This captures the predictability of the current word given a one-word context. For example, the probability of the word *in* given that the previous word was *interested* is higher than the probability of *in* if the last word was *dog*. The *backward transitional probability* is the conditional probability of a word given the next word $P(w_k|w_{k+1})$. This provides an estimate of how predictable the current word is given the next word,

e.g., of how probable it is to see *interested* or *dog* next, given the current word is *in*. A possible interpretation of why material that is further back in the text can benefit the current word and lead to shorter reading times for words with high backward transitional probabilities are preview effects and backward saccades. These corpus results are backed up by results demonstrating the role of forward transitional probabilities in controlled reading experiments (McDonald & Shillcock, 2003a; but see Frisson, Rayner, & Pickering, 2006, who equate transitional probability and Cloze predictability).

It is interesting to note that the forward transitional probability $P(w_k|w_{k-1})$ is a simple form of surprisal, viz., one that takes into account only the previous word w_{k-1}, rather than the whole prefix $w_1 \cdots w_{k-1}$ (see Equation (4)). Another difference is that forward transitional probabilities are estimated using word bigrams, while surprisal is typically estimated using syntactically generated probabilities, based on Equations (5) and (6). We will return to this issue in the context of our discussion of surprisal in the Dundee Corpus in Section 5.

In the current paper, we are interested primarily in syntactic processing effects such as the ones captured by DLT integration cost and surprisal. We therefore need to make sure that these metrics account for variance in the eye-movement recorded that is not captured by the low-level linguistic and oculomotor variables discussed above. Technically, this can be achieved by running hierarchical mixed effect models which include both the low-level and the high-level variables as predictors, as well as partitioning out subject variance. This will be detailed in Section 3.1.2.

3. Experiment 1: Integration Cost

The aim of this experiment is to provide a broad-coverage test of Gibson's DLT by investigating whether integration cost is a significant predictor of eye-tracking measures obtained on a corpus of naturally occurring, contextualized text.

3.1. Method

3.1.1. Data

For our data analysis, we used the English portion of the Dundee Corpus (Kennedy & Pynte, 2005), an English language eye-tracking corpus based on texts from *The Independent* newspaper. The corpus contains 20 texts, each comprising approximately the same number of words, split into 40 five-line screens. The corpus consists of 51,502 tokens[3] and 9,776 types in total. It is annotated with the eye-movement records of 10 English native speakers, who each read the whole corpus, and answered a set of comprehension questions after each text. These eye-tracking data were acquired using a Dr. Boise eye-tracker, which recorded the movements of the right eye with a sampling rate of 1 ms and a spatial accuracy of 0.25 characters. (See Kennedy & Pynte, 2005, for further details on the Dundee Corpus.)

Before carrying out our analyses, we excluded all cases in which the word was the first or last one of the line, and also all cases where the word was followed by a any kind of punctuation. This eliminates wrap-up effects that occur at line breaks or at the end of sentences. Furthermore, we excluded all words that were in a region of four or more adjacent words that had not been fixated, since such regions were either not read by the participant or subject to data loss due to tracking errors. This left us with 385,467 words.

[3]The token number refers to tokens as tokenized in the Dundee Corpus for presentation to the participants, i.e., punctuation marks are attached to the words. If words and punctuation marks are counted separately, then there are a bit more than 56k words in the corpus.

The fixation sequence obtained from the eye-tracking experiments can be analyzed by computing a range of eye-tracking measures (see Rayner, 1998, for an overview). The most commonly used ones are first fixation duration, first pass duration, and total reading time. *First fixation duration* is the length of the first fixation that lands on a region. This measure is often assumed to reflect lexical access, but also oculomotor processes and visual properties of the region. *First pass duration* (also known as *gaze duration*) is the sum of all fixations on a region between first entering the region and first leaving it. This measure is thought to be indicative of early syntactic and semantic processing (as well as lexical access). The *total reading time* of a region is the sum of all fixations on a region, including refixations of the region after it was left. This measure is assumed to be indicative of textual integration processes (as well as lexical and syntactic/semantic processing).

For the regression analyses reported in this article, we only included those words which had a non-zero reading time for a given measure (i.e., only those words that were not skipped). For first fixation duration and first pass duration, we thus had 200,684 data points, and 240,157 data points in the total duration analyses.[4] The reader is referred to the Appendix for details regarding data preprocessing.

3.1.2. Statistical Analysis

The statistical analyses in this paper were carried out using linear mixed effects models (Pinheiro & Bates, 2000). These models can be thought of as a generalization of linear regression that allows the inclusion of random factors (such as participants or items) as well as fixed factors. The fixed factors can be discrete (such as whether the previous word was fixated) or continuous (such as word frequency). More specifically, we used hierarchical linear mixed effects models, which make it possible to partition the variance to be accounted for into a number of levels; participants were entered as a separate level in the model, following Richter's (2006) recommendations for the treatment of reading time data (this is a generalization of an approach initially proposed by Lorch & Myers, 1990; for alternative proposals using mixed models, see Baayen, Davidson, & Bates, 2008).

A separate mixed effects model was computed for each of the three dependent variables (first fixation duration, first pass duration, and total reading time). The following low-level predictor variables were entered into each of the models: word length in characters, log-transformed word frequency, forward transitional probability, backward transitional probability, word position in the sentence, whether the previous word was fixated or not, launch distance, and fixation landing position. In addition, one or more predictor variables were included that represented the target measure, i.e., integration cost or surprisal.

Minimal models were obtained by entering all predictors and all possible binary interactions between them into the model and then simplifying the model using the Akaike Information Criterion (AIC). The AIC is a measure that optimizes model fit by taking into account the amount of variance explained as well as the number of degrees of freedom. This procedure ensures that a model is obtained which achieves the greatest fit to the data with the minimum number of predictor variables. In the remainder of the paper, we will give the coefficients and significance levels for those predictors that remain in the minimal model. All of these coefficients are statistically significant, with the possible exception of main effects, which are only removed from the minimal model if there is no significant interaction that depends on them.

[4]By data point we mean the word reading times according to the relevant measure.

3.1.3. Implementation

Non-syntactic Predictors. The non-syntactic predictors used were word length in characters (WORDLENGTH), word position in the sentence (SENTENCEPOSITION), whether the previous word was fixated (PREVIOUSWORDFIXATED), the distance between the previous fixation and the current fixation (LAUNCHDISTANCE), and the character on which the eye lands in the word (LANDINGPOSITION). These values can be read off directly from the Dundee Corpus. The predictors logarithmic word frequency (WORDFREQUENCY), logarithmic word frequency of the previous word (PREVIOUSWORDFREQUENCY), forward transitional probability (FORWARDTRANSITIONALPROBABILITY), and backward transitional probability (BACKWARDTRANSITIONALPROBABILITY) need to be estimated from a training corpus. We used the British National Corpus (BNC) (Burnard, 1995) and estimated unigram and bigram probabilities using the CMU-Cambridge Language Modeling Toolkit (Clarkson & Rosenfeld, 1997). For the bigram model, many of the bigrams from the Dundee Corpus were not observed in the BNC training data. To avoid having to assign a bigram zero probability just because it did not occur in the training data, we smoothed the bigram probabilities, i.e., some of the probability mass of the seen events was redistributed to unseen events. We used the Witten-Bell smoothing method (Witten & Bell, 1991), which is predefined in the CMU Toolkit.

Integration Cost. It is not feasible to manually compute values for the predictor integration cost (INTEGRATIONCOST) for the whole Dundee Corpus, given its size. We therefore relied on automatic methods which can handle a large amount of data (but are potentially error-prone). We parsed the corpus with an automatic parser and implemented a function that uses these parses to assign integration cost values to the words in the corpus. The parser used was Minipar (Lin, 1998), a broad-coverage dependency parser for English. Minipar is efficient and has good accuracy: an evaluation with the SUSANNE corpus (Sampson, 1995) shows that it achieves about 89% precision and 79% recall on dependencies (Lin, 1998) on SUSANNE. A dependency parser was chosen because the dependency relationships that it returns are exactly what we need to calculate integration costs (see Figure 1 for an example).

In our implementation, integration costs are composed of the cost of (a) constructing a discourse referent and (b) the number of discourse referents that occur between a head and its dependent, excluding the head and the dependent themselves. This requires discourse referents to be identified in the corpus; we used the approximation that all words that have a nominal or verbal part of speech are discourse referents. Using part of speech tags assigned by the parser also allows us to differentiate between auxiliaries, modals and full verbs, and to automatically identify nouns that are parts of compound nouns.

It is important to note that two versions of integration cost exist in the literature: one based on Gibson's (2000) DLT, and the earlier version based on Gibson's (1998) syntactic prediction locality theory, a predecessor of DLT. The difference between the two versions only concerns nouns; in this paper, we assume the Gibson (2000) version of integration cost (though we conducted some experiments with the 1998 version, see Section 4.3). DLT has later been extended and revised to provide a more extensive account of noun phrases (e.g., Warren & Gibson, 2002), but this revised version of DLT has not been formalized, and thus would be hard to implement without making additional assumptions.

We evaluated our integration cost implementation using a short text that had been hand-annotated with integration cost values. This evaluation gives us an estimate of how well our auto-

matic annotation tool performs. We found that the integration cost values assigned automatically to the 764 words in the evaluation text were correct 83% of the time. Further analysis revealed that the automatically assigned integration cost values were significantly correlated with the manually assigned ones (Pearson's $r = 0.697$, $p < 0.001$). This result needs to be regarded as a lower bound. Unlike the Dundee Corpus, the evaluation text was not a newspaper text. Rather, it was a manually constructed story created in order to contain sentences with high integration cost. The sentences in the evaluation text are often long and complicated, and therefore hard to analyze with our automatic tool. Mean integration cost in the evaluation text was 0.7, while in the Dundee Corpus it was 0.55.

3.2. Results

In Experiments 1 and 2, we will only consider results for first pass durations in detail. The results for first fixation durations and total times are broadly similar, and will only be discussed briefly. We will return to this in Experiment 3, which provides a comparison of the results for the three eye-tracking measures for a model that contains all the predictors used in this paper (see Section 5.3).

Tables 1 and 2 show the coefficients and significance levels obtained when running hierarchical linear mixed effects models on first pass durations extracted from the Dundee Corpus. Both models include all the non-syntactic predictors and integration cost, and were computed over all words in the corpus. The difference between them is that in Table 1, all predictors were included as main effects only, i.e., no interactions between predictors were included. The interactions between predictors also have explanatory power, but it is informative to first consider a mixed effect model without these interactions. We use this simpler model to explain how to interpret mixed effects models; many of the previously established findings in the reading literature are confirmed by our data. Table 1 shows an intercept of approximately 275 ms. This can be regarded as the base reading time of a word, to which the value for each predictor multiplied by the coefficient for that predictor is added to obtain the predicted reading time for that word.

For example, the coefficient of WORDLENGTH is approximately 15 ms, which means that for each letter of the word, an additional 15 ms are added to the word's predicted reading time. The fact that the coefficient of WORDLENGTH is positive means that longer words have longer reading times, a basic finding in the reading literature. We also observed a negative coefficient for logarithmic word frequency (WORDFREQUENCY), which means that more frequent words are read faster than less frequent words.

We also find that the presence of a fixation on the previous word (PREVIOUSWORDFIXATED) reduces reading time by 25 ms, i.e., fixation time is longer when the previous word was skipped. There is also an effect of landing position (LANDINGPOSITION), whose negative coefficient indicates that reading time decreases with increasing landing positions, at a rate of approximately 10 ms per character. It has been claimed that readers speed up while they move through a sentence (Ferreira & Henderson, 1993). Our data support this, since we obtain a small negative coefficient for the position of the word within the sentence (SENTENCEPOSITION), which means later words are read faster. There was no significant effect of launch distance (LAUNCHDISTANCE), which probably indicates that any variation in reading time due to launch distance is already explained by PREVIOUSWORDFIXATED and LANDINGPOSITION. (Recall that non-significant predictors are removed by our model fitting procedure, that is why LAUNCHDISTANCE does not appear in Table 1.)

For forward transitional probability (FORWARDTRANSITIONALPROBABILITY), we observed a negative coefficient, which means that words with high transitional probability are read

Predictor	Coefficient	Significance
(INTERCEPT)	275.25	***
WORDLENGTH	14.69	***
WORDFREQUENCY	−12.16	***
PREVIOUSWORDFREQUENCY	−5.76	***
PREVIOUSWORDFIXATED	−24.65	***
LANDINGPOSITION	−9.99	***
SENTENCEPOSITION	−0.23	***
FORWARDTRANSITIONALPROBABILITY	−0.54	*
BACKWARDTRANSITIONALPROBABILITY	3.41	***
INTEGRATIONCOST	−2.28	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 1: First pass durations for all words in the Dundee Corpus: coefficients and their significance levels for a model that includes all predictors only as main effects.

faster, in line with McDonald and Shillcock's (2003b) results. However, while McDonald and Shillcock (2003b) also find a negative coefficient for backward transitional probability, while in our data BACKWARDTRANSITIONALPROBABILITY shows a small positive coefficient, which means that words with higher backwards transitional probability show slightly higher reading times.

While the coefficients for the non-syntactic predictors have plausible interpretations that are consistent with the previous literature on reading, the result for the integration cost predictor (INTEGRATIONCOST) is disappointing: we obtained a significant negative coefficient, which means that higher integration cost leads to shorter reading time, contrary to the prediction of DLT.

The same significant predictors were obtain we ran mixed effect models for first fixation duration and in total reading times (we omit the tables here), with one exception: for first fixations, there was no effect of word length and no effect of integration cost.

One potential explanation for the lack of an effect of integration cost may be the fact that (following Gibson), we assumed identity as our integration cost function, i.e., I(n) = n. It is possible that there is a logarithmic relationship between integration cost and reading time (e.g., similar to that between frequency and reading time). We tested this by re-running the analysis reported in Table 1 with the integration cost function I(n) = log(n+1). However, again a significant negative coefficient for INTEGRATIONCOST was obtained (though model fit improved slightly).

We now return to Table 2, which lists the results for a mixed effects model that includes all predictors as main effects and all binary interactions between predictors, and was optimized by removing all predictors that do not improve model fit (see Section 3.1.2). The results are broadly similar to those obtained using main effects only, with the exception that launch distances is now a significant, negative predictor. However, we find significant interaction in this model which makes the coefficients harder to interpret. For example, contrary to expectation, frequency now has a positive coefficient. This needs to be interpreted in the context of the negative coefficient of WORDLENGTH:WORDFREQUENCY, the interaction between word length and frequency.

This interaction means that short, frequent words have longer reading times (positive coefficient of WORDFREQUENCY) than less frequent words with equal length.[5] Examples include

[5]More precisely, the coefficient of frequency becomes negative for words with two letters or more, as $c_f + 2c_{lf} < 0$,

Predictor	Coefficient	Significance
(INTERCEPT)	168.06	***
WORDLENGTH	29.64	***
WORDFREQUENCY	7.54	***
PREVIOUSWORDFREQUENCY	−5.67	***
PREVIOUSWORDFIXATED	−25.62	***
LANDINGPOSITION	1.92	***
LAUNCHDISTANCE	−1.35	***
SENTENCEPOSITION	−0.21	***
FORWARDTRANSITIONALPROBABILITY	−2.00	***
BACKWARDTRANSITIONALPROBABILITY	2.14	***
INTEGRATIONCOST	−2.01	***
WORDLENGTH:WORDFREQUENCY	−3.87	***
WORDLENGTH:LANDINGPOSITION	−1.71	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 2: First pass durations for all words in the Dundee Corpus: coefficients and their significance levels for a model that includes all predictors as main effects and all binary interactions, minimized using the AIC.

abbreviations, or expressions such a *$5*. Among longer words, more frequent ones are read faster, as expected (negative coefficient of WORDLENGTH:WORDFREQUENCY). Similarly, we observe a significant negative coefficient for the interaction of word length and landing position. The interpretation is analogous to that of the WORDLENGTH:WORDFREQUENCY interaction: the positive effect of landing position on reading time is reversed for longer words.

Crucially, the coefficient for integration cost is negative also in the model that includes all predictors an all binary interactions. Again, this runs against the DLT prediction that higher integration cost should lead to higher reading times.

When we fitted mixed models for first fixation times and total times, we again found the same pattern of results as for first pass time, with the exception that the INTEGRATIONCOST effect was not significant in first fixations.

3.3. Discussion

In this experiment, we fitted mixed effect models on the reading times for all words in the Dundee Corpus, and found that integration cost is a significant negative predictor of reading time, i.e., that higher integration cost values correspond to shorter reading times, contrary to the prediction of DLT. This result can be explained by the fact that DLT only provides a partial definition of syntactic processing complexity: integration costs are only assigned to nouns and verbs. All other words have an integration cost of zero, while there are very few nouns or verbs with an integration cost of zero (only non-head nouns in compounds).

We therefore further investigated the relationship between reading time and integration cost. We re-ran the mixed effects model in Table 2 on all words in the corpus and included integration cost as a factor, i.e., as a discrete predictor. When the DLT predictions are entered into the regression as categorical values, separate coefficients are estimated for each integration cost value.

where c_f and c_{lf} are the coefficients of WORDFREQUENCY and WORDLENGTH:WORDFREQUENCY, respectively.

These separate coefficients allow us to assess the influence of words with an integration cost of zero: the negative overall coefficient for integration cost as a continuous variable may be due to the fact that words with integration cost 0 are problematic, because not all of them may be covered by the theory. Therefore it is interesting to see whether there is an overall positive trend for words that are assigned an integration cost. Figure 3 plots integration cost values against their model coefficients and shows a general trend of higher integration cost values corresponding to greater coefficients (i.e., increased reading times), as predicted by DLT. The figure also shows that the coefficients for integration cost values one to nine are negative, i.e., the reading times for words with these integration cost values is shorter than the reading time for words with zero integration cost (which the model takes as the base value and assigns a coefficient of zero). This finding indicates that words with integration cost 0 can still generate difficulty, but that this difficulty is not captured by DLT, which only makes predictions for nouns and verbs. This result also means that the current coverage of DLT is clearly not sufficient for naturally occurring text. Most words in the corpus have integration cost values between zero and nine.[6] This explains why we found an overall negative coefficient of integration cost in Table 2 (where INTEGRATIONCOST was entered as a continuous predictor), even though higher integration cost values generally correspond to higher reading times in Figure 3.

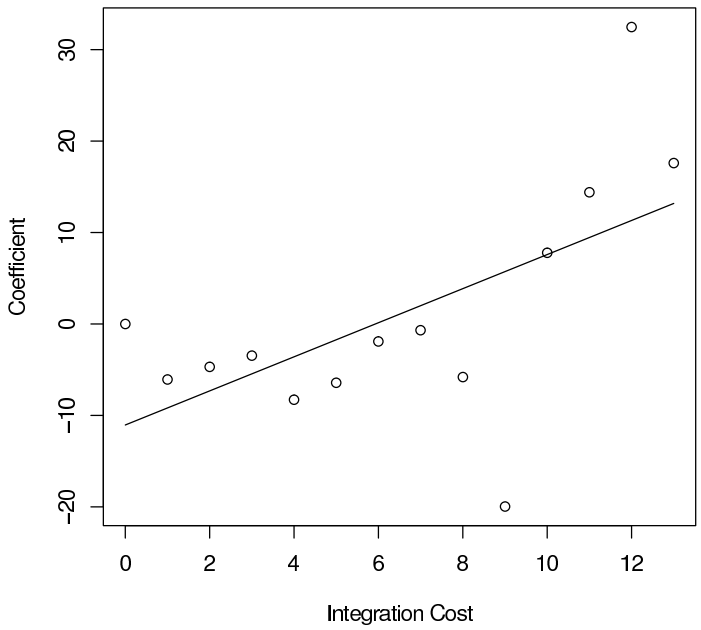


Figure 3. Coefficients for the factor integration cost in a mixed effects model on the words in the Dundee Corpus.

As Figure 3 shows, the average residual reading time of words with zero integration cost is higher than those of words with slightly higher integration cost. Since DLT traditionally only makes predictions for verbs and nouns, it would be interesting to find out at what other word types a similar cost might be incurred. To test whether some types of words take longer to read than others

[6] In fact, the largest influence on the regression coefficient comes from words with integration cost 0 (approx. 125,000 fixated words) and integration cost 1 (approx. 84,000 fixated words).

after factoring out low level effects, we computed residual reading times on the Dundee Corpus by building a mixed effects model that contains all the non-syntactic predictors, and subtracted the reading times predicted by this model for the observed reading times. We analyzed these data by partitioning them according to the words' parts of speech (POS). We found that adjectives, prepositions, sentence adjectives, and expletives have mean residual reading times larger than zero, which means they are read slower than would be expected according to word length, frequency, and the other non-syntactic predictors. The data suggests that it could be interesting to extend DLT in a way that makes it possible to also assign an integration cost to those word categories.

4. Experiment 2: Integration Cost for Verbs and Nouns

In Experiment 1, we obtained a negative coefficient for integration cost when we fitted a mixed effects model to predict reading times for all words in the Dundee Corpus. We concluded that this finding is due to the fact that DLT does not make integration cost predictions for words other than verbs and nouns. In the present experiment, we will explore this link further by providing a detailed analysis of integration costs for nouns and verbs.

4.1. Method

Data, statistical analysis, and implementation used were the same as in Experiment 1.

4.2. Results

Again, we will only consider results for first pass durations in detail; the reader is referred to Experiment 3 (see Section 5.3) for a more detailed comparison of results for first fixation durations, first past times, and total times.

Nouns. We first fitted a mixed effects model for the first pass durations for all the nouns in the Dundee Corpus (49,761 data points for the early measures, 57,569 data points for total durations) that included all predictors as main effects and all binary interactions, minimized using the AIC. Integration cost was not a significant, positive predictor of reading time in this model.

When the data set was restricted further, viz., to nouns with non-zero integration cost (45,038 and 51,613 data points respectively), a significant, positive coefficient for integration cost was obtained. Furthermore, we found that model fit improves slightly when using the logarithmic integration cost function I(n) = log(n + 1) compared to when using a linear one. The coefficients of this model are listed in Table 3. The significant positive coefficient for integration cost in this model means that nouns with higher integration cost take longer to read.

We fitted mixed models for first fixation durations and total times, and found the same set of significant predictors, with the following exceptions: for first fixations, there was no significant effect of WORDLENGTH, and the effect of INTEGRATIONCOST was small, and there were no significant interactions. For total times, INTEGRATIONCOST narrowly failed to reach significance ($p = 0.07$).

We further investigated why the effect of integration cost on nouns was only present if nouns with zero integration cost were excluded. This is particularly puzzling as it is rare that nouns receive an integration cost of zero; there is only way for this to happen in the corpus: the first word of noun-noun compounds and pronouns. We re-ran the model in Table 3, but included pronouns (an additional 4,840 data points for the early measures, 6,108 data points for total durations), despite

Predictor	Coefficient	Significance
(INTERCEPT)	128.24	***
WORDLENGTH	30.90	***
WORDFREQUENCY	14.50	***
PREVIOUSWORDFIXATED	−18.05	***
LANDINGPOSITION	−4.18	***
LAUNCHDISTANCE	−1.91	***
SENTENCEPOSITION	−0.12	*
FORWARDTRANSITIONALPROBABILITY	−3.27	***
BACKWARDTRANSITIONALPROBABILITY	3.96	***
log(INTEGRATIONCOST)	5.86	*
WORDLENGTH:WORDFREQUENCY	−4.98	***
WORDLENGTH:LANDINGPOSITION	−1.02	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 3: First pass durations for nouns (with non-zero integration cost) in the Dundee Corpus: coefficients and their significance levels for a model that includes all predictors as main effects and all binary interaction, minimized using AIC.

their integration cost of zero. Again, a significant, positive coefficient of integration cost was obtained. First parts of compounds were relatively frequent in the Dundee corpus: there were 7,121 data points for total durations and 6,118 data points for the early measures; a large proportion of these cases consisted of proper names (such people's names or titles).

Verbs. Just as for nouns, we fitted a mixed effects model for the first pass durations for all the verbs in the Dundee Corpus (the model again included all main effects and all binary interactions). No significant, positive coefficient for integration cost was obtained in this model. We re-ran the model with verbs that exhibit a non-zero integration cost, and with a logarithmic instead of a linear integration cost function. Again, integration cost was not a significant, positive predictor of reading time.

We then fitted a model that included the part of speech of the verb as a predictor. The rationale behind this is that verb reading time differs by part of speech, e.g., inflected verbs are read more slowly than infinitives. This model only included verbs with non-zero integration costs and used a logarithmic integration cost function. We found that integration cost was a significant, positive predictor of reading time (though the size of the coefficient was smaller than for nouns).

In order to further investigate the integration cost effect that we found for verbs, we computed residual reading times for this data set (see Section 3.3). On the residuals, we then fitted a model that includes a predictor that indicates the part of speech of the dependent that is integrated at a given verb (or sequence of parts of speech if multiple dependents are integrated). The coefficients in this model indicate which dependents lead to higher or lower integration costs, see Table 4. We observe that, as predicted by DLT, the integration of nouns (parts of speech NN, NNP, NNS) or adverbs (part of speech RB) leads to longer reading times, unless there is also an auxiliary (AUX) that occurs before the verb. The auxiliary thus seems to facilitate integration of nouns at the verb.

Dependents	Coefficient	Significance	*N*
PRP-AUX-NN	−81.45	**	15
PRP-AUX	−76.24	**	13
NNP-AUX-AUX	−62.41	**	21
RP	−62.34	*	12
NNP-AUX	−59.52	*	17
PRP-MD	−56.44	*	17
NNS-AUX-AUX	−35.65	*	57
NNS-MD-AUX	−30.75	**	110
PRP-AUX-PR-PAUX	−29.72	***	184
NN-MD-AUX	−25.35	**	153
PRP-AUX	−22.64	***	700
PRP-AUX-RB	−21.75	*	133
AUXG	−20.26	*	121
NNP-AUX	−19.05	**	301
TO-PRP	−16.97	***	723
NNP	12.01	**	1372
NN-RB	22.26	*	127
AUX-NNP	66.11	*	15
VBP	67.69	*	10
RB	75.88	**	15
NN-NNS	76.43	***	25
PRP-MD-PRP-MD-JJ	105.4	*	65

Table 4: First pass durations for verbs (with non-zero integration cost) in the Dundee Corpus: coefficients for the verbal dependents and their significance levels for a model fitted on residual reading times. Abbreviations in the table refer to part of speech tags used by the Penn Treebank annotation: AUX: auxiliary, PRP: personal pronoun, NN: singular or mass noun, NNP: proper noun, singular, RP: particle, MD: modal, NNS: plural noun, RB: adverb, AUXG: auxiliary present participle, TO: preposition *to*, JJ: adjective, VBP: non-third person singular present verb.

4.3. Discussion

In Experiment 1, we saw that DLT integration cost does not constitute a broad-coverage theory of syntactic complexity, in the sense that integration cost failed to emerge as a significant, positive predictor of reading time on the whole of the Dundee Corpus. We hypothesized that this is due to the fact that DLT only makes partial integration cost predictions, viz., for nouns and verbs only. In the present experiment, we investigated this further by analyzing the performance of DLT on verbs and nouns in more detail.

We showed that integration cost is a significant, positive predictor of reading time on nouns with a non-zero integration cost, and thus supports the hypothesis in DLT. However, this result reflects only effects on a small amount of the data: In its standard form (Gibson, 2000), DLT does not make very interesting predictions for nouns. By default, all nouns have an integration cost of one, because a discourse referent is built. The only cases in which nouns can receive an integration cost of greater than one are in constructions such as *request for permission*, where *permission* is analyzed as the head of the NP, genitive constructions like *the Nation's criminals*, and copula constructions.

In the latter, nouns are considered to be the head of the phrase and integrate the verb *be*. This means that the integration cost for the noun depends on the number of discourse referents intervening between the noun and *be*.

We also investigated the two cases in which DLT assigns an integration cost of zero to nouns. The first case is pronouns, which DLT assumes to constitute old discourse referents, not incurring a cost. We extended our model by including pronouns (as the only nouns with zero integration cost), and still found that integration cost was a significant, positive predictor, which provides evidence for the DLT assumption that pronouns carry zero integration cost. The second case of zero integration cost is noun-noun compounds, for which DLT assume that the first noun incurs no integration cost. However, when we fitted a model on all nouns (including the ones with zero integration cost), we failed to obtain a significant coefficient for integration cost. This indicates that the DLT assumption of cost-freeness for the first noun of a noun-noun compounds is incorrect. Rather, we have to assume that a discourse referent is already being established when the first noun in the compound is encountered, i.e., this noun should incurs a non-zero cost.

At this point, it becomes important which version of DLT is used to compute integration cost values. In contrast to the Gibson (2000) version used in this paper, the Gibson (1998) version of DLT assigns higher integration costs to nouns that occur after their head noun. In order to test how crucial this assumption is, we implemented the 1998 version and conducted the same experiments as with the 2000 version, but this failed to yield an improved fit on our data set.

In addition to looking at nouns, we also investigated the relationship between reading times and integration cost for verbs and were able to show that integration cost is a significant positive predictor of verb reading times. However, the coefficient was small compared to that found for nouns; also, this result was only obtained for a model that includes the parts of speech of the verbs as an additional predictor. This indicates that integration cost only has a small overall effect on reading time for verbs, and that this effect is variable across parts of speech.

As verb integration cost is at the heart of DLT (which predicts only limited variation in noun integration cost, see above), we investigated this result further. We fitted a model on the residual reading times that included the parts of speech of the dependents to be integrated at the verb as a predictor. This analysis revealed the following pattern (see Table 4): positive coefficients were obtained for the integration of nominal dependents (indicating that this integration leads to increase reading time), while negative coefficients were obtained for the integration of auxiliaries (which means that this integration decreases reading time). In this context, it is interesting to note that Warren and Gibson (2002) found a reading time effect for auxiliaries. Auxiliaries following definite NPs were read more slowly than auxiliaries following pronouns. This result in consistent with our findings in the Dundee Corpus, i.e., that auxiliaries, and not just main verbs, show integration cost effects. However, Warren and Gibson (2002) interpret their finding as a spillover effect.

5. Experiment 3: Surprisal

Experiments 1 and 2 indicate that there is evidence that DLT integration cost is a predictor of reading time in the Dundee Corpus. However, DLT cannot be regarded as a broad coverage model, as we found integration cost effects only if we limited our models to verbs and nouns with non-zero reading times. The present experiment has the aim of evaluating surprisal as an alternative model of syntactic processing complexity. Unlike DLT, surprisal is designed to make predictions for all words in a corpus, on the basis of a probabilistic grammar. We will test two versions of surprisal (lexicalized and unlexicalized), and compare them against non-syntactic probabilistic predictors of

reading time (forward and backward transitional probability). Finally, we will also investigate a possible relationship between surprisal and integration cost.

5.1. Method

Data and statistical analysis were the same as in Experiments 1 and 2. For calculating the surprisal values for the words in our corpus, we parsed the Dundee Corpus with an incremental parser which returns a prefix probability for each word in the corpus, i.e., the probability in Equation (5). We can then use Equation (6) to obtain the surprisal value for a word w_{k+1}: we subtract the logarithmic prefix probability for w_{k+1} from the logarithmic prefix probability for w_k. The parser used was Roark's (2001) incremental top-down parser. This is a probabilistic parser trained on the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1993), a corpus of English text manually annotated with phrase structure trees. Only the Wall Street Journal section of the Penn Treebank was used for training. The parser achieves a broad coverage of English text and is highly accurate, with a precision and recall of 85.7% for labeled brackets reported by Roark (2001). As the Dundee Corpus also consists of newspaper text comparable to the Wall Street Journal text the parser was trained on, we can expect a similar performance on the Dundee Corpus.

We estimated surprisal in two different ways. The first version was fully lexicalized, i.e., it takes into account the exact words of a string when calculating structural and lexical probabilities. This lexicalized version was obtained using the default configuration of the Roark parser. The second version was unlexicalized, i.e., only used the structural probabilities. The unlexicalized model does not take into account word frequency or the probability of a word being assigned a specific POS tag (i.e., there are no lexical rules of type $V \rightarrow$ *wrote*). This structural version of surprisal helps us to factor out frequency effects, but is also limiting in that no subcategorization information is available to the model for calculating structural probabilities, as this information is contained in the lexical rules. To use the Roark parser for calculating an unlexicalized version of surprisal, we replaced each word by its own part-of-speech tag and trained the parser on the POS tag sequences. This eliminates the effect of word frequencies.

5.2. Results

Table 5 shows the coefficients and significance levels obtained when running a mixed effects model on first pass durations in the Dundee Corpus. As in Experiment 1, this model was computed over all words in the corpus, and included all non-linguistic predictors as well as lexicalized surprisal (LEXICALIZEDSURPRISAL), unlexicalized surprisal (UNLEXICALIZEDSURPRISAL), and forward and backward transitional probability

Table 5 shows unlexicalized surprisal is a significant, positive predictors of reading time (high surprisal leads to longer reading time). The coefficient for UNLEXICALIZEDSURPRISAL is small, but this has to be interpreted in the context of the range of this predictor: the values for unlexicalized surprisal range from 0.04 to 18.1, with a mean surprisal of 2.45.

Lexicalized surprisal (LEXICALIZEDSURPRISAL) does not figure in Table 5, which means that it was not a significant predictor of reading time, and was eliminated from the model during model selection. However, forward transitional probability was a significant negative predictor of reading time (higher probability means lower reading time), and backward transitional probability has a positive coefficient. As detailed in Section 2.3, forward transitional probability can be regarded as a simple form of surprisal that only takes into account the immediate context (the preceding word). This indicates that lexicalized surprisal does not explain any variance in the eye-movement

Predictor	Coefficient	Significance
(INTERCEPT)	135.67	***
WORDLENGTH	29.77	***
WORDFREQUENCY	8.57	***
PREVIOUSWORDFIXATED	−17.70	***
LANDINGPOSITION	1.13	**
LAUNCHDISTANCE	−1.63	***
SENTENCEPOSITION	−0.20	***
FORWARDTRANSITIONALPROBABILITY	−1.60	***
BACKWARDTRANSITIONALPROBABILITY	2.06	***
UNLEXICALIZEDSURPRISAL	1.03	***
WORDLENGTH:WORDFREQUENCY	−4.01	***
WORDLENGTH:LANDINGPOSITION	−1.66	***

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 5: First pass durations for all words in the Dundee Corpus: coefficients and their significance levels for a model that includes all predictors as main effects, and all binary interaction, minimized using the AIC.

record over and above what is explained by forward transitional probability and unlexicalized surprisal.

We also fitted mixed effect models for first fixation durations and total times, which also showed an effect of unlexicalized surprisal, and the absence of lexicalized surprisal. Also the other significant factors listed in Table 5 were significant for first fixations and total times, except for fact that the interaction of WORDLENGTH and LANDINGPOSITION was not significant for first fixations; also all effect sizes were much smaller for this measure.

5.3. Discussion

This experiment showed that surprisal can function as a broad-coverage model of syntactic processing complexity: we found that unlexicalized surprisal was a significant, positive predictors of reading time on arbitrary words in the Dundee Corpus. This sets surprisal apart from integration cost, which does not make predictions for all words in the corpus, and for which we only obtained significant effects on verbs and nouns.

We failed to find a corresponding effect for lexicalized surprisal. This indicates that means that forward transitional probability and structural surprisal taken together are better predictors of reading times in the Dundee Corpus than lexicalized surprisal, which combines these two components. Forward transitional probability can be regarded as a simple approximation of surprisal (see Section 2.3), and our results indicate that this approximation is sufficient, at least when it comes to predicting the reading times in the corpus.

Unlexicalized surprisal, on the other hand, takes structural probabilities into account, but disregards lexical probabilities, and therefore is a significant predictor of reading time, even if forward transitional probability is also entered into the model. We conclude that structural surprisal is able to explain a component in the reading time data that neither lexicalized surprisal, nor transitional probabilities, nor any of the other non-syntactic predictors can explain. This is evidence for Hale's (2001) and Levy's (2008) hypothesis that the incremental disconfirmation of syntactic hypotheses by the parser can explain processing complexity.

This raises the more general question of overlap between the various measures of syntactic processing complexity investigated in this paper. To address this issue, we computed correlations between integration cost and the different incarnations of surprisal (lexicalized and unlexicalized surprisal, forward and backward transitional probabilities), and word frequency. The result is given in Table 6; all correlations are statistically significant except for the pair WORDFREQUENCY–UNLEXICALIZEDSURPRISAL (even small correlations are significant due to the large number of observations). As expected, forward and backward transitional probability are highly correlated. Furthermore, the lexicalized measures (lexicalized surprisal and transitional probabilities) are highly correlated with word frequency. The high correlation between lexicalized surprisal and forward transitional probability confirms the intuition that these two measures are in fact both incarnations of surprisal, but of a different level of granularity. On the other hand, structural surprisal is not significantly correlated with the other measures, including word frequency (though there is a weak correlation with lexicalized surprisal). This confirms that unlexicalized surprisal really captures structural probability effects, without taking lexical probabilities into account. Crucially, Table 6 also shows that integration cost is orthogonal to surprisal and the other frequency-based predictors: there is no strong correlation between INTEGRATIONCOST and any of the other predictors. This is supporting evidence for our hypothesis that both DLT and surprisal capture relevant aspects of processing difficulty, but that these aspects are complementary, since DLT describes difficulty incurred through memory load effects and reactivating previous material to integrate it into the current context, whereas surprisal captures the predictability of the context and changes in the maintained interpretations.

This finding holds even if we compute correlations only for the verbs in the Dundee Corpus (not shown in the table): the correlation between integration cost and unlexicalized surprisal is approximately 0.05 for verbs, while the correlation between integration cost and lexicalized surprisal is approximately 0.01 for verbs. This confirms that integration cost and surprisal are orthogonal: if there was a relationship between them, it should manifest itself on verbs, as verbs are the words with the largest variation in integration cost (compared to nouns, which mostly have an integration cost of one, and the other words in the corpus, which have an integration cost of zero; see also Section 4.3).

Finally, we fitted a mixed effects model that includes lexicalized and unlexicalized surprisal, forward and backward transitional probability, as well as integration cost. To illustrate the differences between various eye-movement measures, we fitted separate models for first pass duration (the measure discussed so far), and additionally first fixation time, and total time.[7] The results are given in Table 7. We will first discuss first pass times, which showed that integration cost, unlexicalized surprisal, lexicalized surprisal, as well as forward and backward transitional probability are all significant predictors of reading time. However, the coefficient of integration cost was negative, confirming that integration cost is not a broad-coverage predictor of reading time (as shown in Experiment 1). Furthermore, LEXICALIZEDSURPRISAL, while significant, has a small negative coefficient, meaning that words with higher lexicalized surprisal show longer reading times. This is compatible with the model in Table 5, which failed to find a significant effect of unlexicalized surprisal.

Turning to the results for first fixation times (see Table 7), we again found a significant neg-

[7]Note that these models are based on different subsets of the data, since the data sets include all words that have non-zero reading time. This means that total times have more data points (in addition to first pass and first fixations all those that were not fixated in first pass, but at some later pass.

	INTEGR COST	WORD FREQ	LEX SURPRIS	UNLEX SURPRIS	FORWTRANS PROB
WORDFREQUENCY	−0.25				
LEXSURPRISAL	0.17	−0.57			
UNLEXSURPRISAL	−0.07	0.04	0.36		
FORWTRANSPROB	−0.20	0.62	−0.66	−0.10	
BACKTRANSPROB	−0.26	0.62	−0.53	0.04	0.68

Table 6: Correlation coefficients (Pearson's r) between the predictors, for fixated words ($N = 237,163$).

ative effect of forward transitional probability, and a significant positive effect of backward transitional probability. Unlexicalized surprisal was a positive predictor of reading time, while integration cost and lexicalized surprisal were removed by the model selection procedure because they were non-significant. As in the previous experiments, the coefficients for first fixation times were smaller than the ones for first pass times.

The results for total time (see also Table 7) replicated the results for first pass; again forward and backward transitional probability, integration cost, and lexicalized and unlexicalized surprisal were significant predictors. The coefficients for integration cost and surprisal were negative, also replicating the findings for first pass times.

	First Fix		First Pass		Total Time	
Predictor	Coef	Sig	Coef	Sig	Coef	Sig
(INTERCEPT)	193.25	***	143.17	***	196.15	***
WORDLENGTH	1.74	***	29.54	***	23.73	***
WORDFREQUENCY	−2.57	***	7.05	***	4.49	***
PREVIOUSWORDFIXATED	−6.42	***	−17.72	***	−27.51	***
LANDINGPOSITION	rem	–	1.23	**	n/a	–
LAUNCHDISTANCE	−1.81	***	−1.62	***	n/a	–
SENTENCEPOSITION	−0.05	***	−0.20	***	−0.26	***
FORWARDTRANSITIONALPROBABILITY	−2.06	***	−2.14	***	−2.45	***
BACKWARDTRANSITIONALPROBABILITY	0.45	**	1.55	***	1.55	**
log(INTEGRATIONCOST)	rem	–	−5.37	***	−6.99	***
LEXICALIZEDSURPRISAL	rem	–	−0.73	***	−1.16	***
UNLEXICALIZEDSURPRISAL	0.39	***	1.39	***	2.18	***
WORDLENGTH:WORDFREQUENCY	−0.47	***	−3.86	***	−4.15	***
WORDLENGTH:LANDINGPOSITION	rem	–	−1.67	***	0.13	–

$*p < 0.05$, $**p < 0.01$, $***p < 0.001$

Table 7: First fixation times, first pass durations, and total time for all words in the Dundee Corpus coefficients and their significance levels for a model that includes both surprisal and integration cost as predictors, minimized using the AIC. Predictors marked "n/a" are not applicable for this measure; Predictors marked "rem" were removed from the regression because they did not significantly reduce the AIC.

6. General Discussion

In this paper, we evaluated two theories of syntactic processing complexity against reading time data extracted from a large eye-tracking corpus: Gibson's (1998, 2000) Dependency Locality Theory (DLT) and Hale's (2001) surprisal. We selected these two approaches for our investigation because they make complementary theoretical assumptions: DLT's integration cost captures the cost incurred when a head has to be integrated with the dependents that precede it. Surprisal, on the other hand, accounts for the cost that results when the current word is not predicted by the preceding context.

This paper evaluated integration cost and surprisal using a broad coverage approach, i.e., we investigated whether the two theories provide accurate predictions for arbitrary words in naturalistic, contextualized text (as opposed to artificially constructed experimental materials, presented out of context and repeated many times). For this investigation we used the reading time data in the Dundee corpus, a large corpus of newspaper text annotated with eye-movement data.

We found that DLT's integration cost was not able to provide reading time predictions for the Dundee corpus as a whole. This was largely due to the fact that DLT only assigns integration cost values to verbs and nouns; this means that the majority of words in the corpus have an integration cost of zero. However, we were able to show that integration cost is a significant predictor of reading time if the verbs and nouns in the corpus are analyzed separately. We also identified limitations of DLT's treatment of nouns. One example is the assumption that the first noun in noun-noun compounds carries zero integration cost. This is incompatible with our results, which indicate that the integration cost should be spread over the whole compound. Furthermore, we observed that DLT only makes a restricted range of predictions for nouns: with few exceptions, all head nouns are assigned an integration cost of one. Arguably, this limits the power of the theory in explaining reading time data for noun phrases in a corpus, which are often complex. This problem could be address by extending DLT along the lines suggested by Warren and Gibson (2002). They provided evidence that processing complexity at the verb varies with the referential properties of the NP to be integrated, as expressed by the NP's position on the Givenness Hierarchy (Gundel, Hedberg, & Zacharski, 1993). They find that complexity increases from pronouns to names to definite NPs to indefinite NPs. Warren and Gibson (2002) suggest that a continuous integration cost metric needs to be developed that takes the givenness status of the integrated NP into account. This would result in a wider range of integration cost values for the nouns in the Dundee Corpus, potentially making it possible to explain more variance in the reading time record.

When we tested DLT predictions against the verbs in the Dundee corpus, we found evidence that the integration cost definition for auxiliaries needs to be revised: verbs that integrate an auxiliary and a nominal dependent exhibit a reduced integration cost compared to verbs that only integrate a nominal dependent. This result has an interesting implication for DLT. On the one hand it confirms the DLT assumption that an integration cost is incurred at the verb when nominal dependents are integrated. On the other hand, it shows that this does not extend to cases where an auxiliary precedes the main verb. A possible explanation is that the relevant integration cost is not incurred at the main verb, but at the auxiliary itself, which integrates nominal dependents and thus incurs a non-zero integration cost (DLT assume that auxiliaries are cost-free). When the auxiliary is then integrated with the main verb, it facilitates integration (hence the negative coefficient), as the main work of the integration of the nominal dependents has already happened at the auxiliary. Note that this assumption is compatible with syntactic theories such as Head-driven Phrase Structure Grammar

(Pollard & Sag, 1994), which assume that auxiliaries inherit the subcategorization frame of the main verb, and that dependents are unified (integrated) into the subcategorization frame at the auxiliary.

At this point, it is worth considering a more radical departure from DLT's assumptions. Integration cost is standardly defined in terms of the number of discourse referents intervening between a head and its dependents, but alternatives have been proposed in the literature. For example, Alexopoulou and Keller (2007) show that two types of extraction from *wh*-phrases can differ in processing complexity, even though they involve the same number of intervening discourse referents. Based on this result, they argue that the number of intervening syntactic heads (rather than discourse referents) is the crucial factor for determining integration cost. This is a hypothesis that could be tested against the Dundee Corpus. A head-based definition of integration cost would results in different complexity predictions for a large number of words in the corpus, possibly resulting in a better fit with the reading time data. We leave this as an issue for future research.

In the second part of this paper, we evaluated the predictions of Hale's (2001) surprisal measure on the Dundee corpus. We computed surprisal in two ways: lexicalized surprisal was estimated using a probabilistic parser that utilizes lexical (word-based) probabilities as well as structural (rule-based) probabilities. Unlexicalized surprisal was estimated using a parser that only has access to structural probabilities. We found that only structural surprisal was a significantly positive predictor of reading times. This finding can be explained by the fact that lexicalized surprisal is highly correlated with word frequency and transitional probability (transitional probability can be seen as a simple approximation of lexicalized surprisal). Therefore, lexicalized surprisal fails to explain any additional variance in the eye-movement record. Unlexicalized surprisal, however, is uncorrelated with word frequency and transitional probability and is able accounts for a part of the variance in reading time that no other predictor captures. This result shows that unlexicalized surprisal is a good candidate for a broad-coverage model of syntactic processing complexity; it generates accurate numerical predictions for all types of words in the corpus, not just for nouns and verbs, as integration cost does.

Our findings regarding lexicalized surprisal indicate that a fully lexicalized parsing model does not offer an advantage over an unlexicalized one. However, this does not mean that there is no role for lexical information in modeling reading times. The experimental literature offers broad evidence for the fact that sentence processing relies on lexical information, such as subcategorization frame frequencies (e.g., Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Trueswell, Tanenhaus, & Kello, 1993) and thematic role preferences (e.g., Garnsey et al., 1997; Pickering, Traxler, & Crocker, 2000). Recent probabilistic models of human sentence processing have attempted to integrate such information with the structural probabilities generated by a parser (Narayanan & Jurafsky, 2002; Padó, 2007). It seems likely that these models (which are effectively unlexicalized model augmented with a limited form of lexical information) would yield a more accurate account of reading times in the Dundee Corpus.

Our surprisal results are corroborated by Ferrara Boston, Hale, Kliegl, Patil, and Vasishth's (2008) work using the Potsdam Sentence Corpus. They found that unlexicalized surprisal is a significant predictor of reading times, even though the Potsdam Sentence Corpus differs in a number of ways from the Dundee corpus: it uses a different language (German) and it consists of unconnected sentences, which were manually constructed for experimental purposes, rather than taken from naturally occurring text. Also, it is smaller in terms of items (144 sentences), but larger in terms of participants (272 participants) than the Dundee corpus. It is therefore encouraging that our results are consistent with Ferrara Boston et al.'s (2008), in spite of these corpus differences. Ferrara

Boston et al. (2008) did not test lexicalized surprisal or integration cost on their data set, but they compared two versions of unlexicalized surprisal, estimated either using a context-free grammar (i.e., in the same way as in the present paper), or using a dependency grammar. In both cases, the surprisal estimates were a significant predictor of reading times.

The analyses reported in this paper were carried out on first-pass reading times computed from the Dundee corpus. We also investigated another early measure (first fixation durations) and a late measure (total times). The results for these two measures are very similar to the ones for first pass, except that first pass showed no effect of integration cost, even if verbs are considered separately. Unlexicalized surprisal, on the other hand, was a significant predictor in all three measures. This finding could indicate that integration cost is associated with later processes in comprehension (that do not manifest themselves in first fixations), while surprisal is associated with both early and late processes (including lexical access, which is often thought to be reflected in first fixation times). This result is corroborated by Ferrara Boston et al. (2008), who also report that unlexicalized surprisal is a significant predictor for all the eye-tracking measures they tested (their analysis involved eight different measures).

Another central finding of the present paper was the fact that surprisal and integration cost are uncorrelated, both for arbitrary words in the corpus, and for verbs (for which DLT makes the bulk of its predictions). This result suggests that a complete theory of sentence processing complexity needs to include two mechanisms: a backward-looking one as proposed by DLT, and a forward-looking one as proposed by surprisal. When a new word is processed it incurs two types of processing cost: the cost of integrating material that has been processed previously with the new word, and the cost of discarding alternative syntactic predictions that are not compatible with the new word. The first type of cost corresponds to locality effects that have been observed extensively in the literature (see Gibson, 1998 for an overview). The second type of cost corresponds to anti-locality effects which have been reported recently (Konieczny, 2000; Vasishth & Lewis, 2006). In order to capture both types of cost (and yield broad-coverage results on an eye-tracking corpus), we need to develop a unified model that encompasses both the prediction of upcoming material and the subsequent verification and integration processes (for a first step towards such a model see Demberg & Keller, 2008).

Another point to consider is the fact that the predictions of both DLT and surprisal depend on the grammar formalism that they are operating on. In DLT, syntactic structures (head–dependent relations) determine the amount of integration cost that is incurred by a given sequence of words. While there are many clear cases of what constitutes the head, the dependent and the relation between them can be subject to debate in the linguistic literature. In the current paper, we assumed that the dependency structures output by Minipar form the basis of the integration cost computations (see Section 3.1.3). Minipar uses one particular codification of dependency grammar (Sampson, 1995), and it is therefore conceivable that our results would change if we computed integration cost using a parser that makes a different set of representational assumptions.

It is important to note that surprisal also requires representational assumptions. The definition of surprisal in Equation (4) does not mention syntactic structures explicitly. However, in order to compute the conditional probability in this equation, prefix probabilities have to be obtained, which requires summing over all possible analyses of a string. The number and type of these analyses will differ between grammatical frameworks, which entails that representational assumptions do play a role for surprisal. In the present paper, we only investigated the predictions of one type of syntactic representations, viz., those of Roark's (2001) parser, which generates Penn Treebank-style

structures. It is possible that other syntactic models will yield different surprisal estimates and fit the reading time data more closely, or model different aspects of the data. (This has been investigated by Ferrara Boston et al., 2008, who compare dependency and phrase-structure versions of surprisal, as detailed above.)

Apart from its theoretical contribution, this paper also makes a methodological contribution. To our knowledge, this is the first time that theories of sentence processing have been tested on broad-coverage data extracted from an eye-tracking corpus.[8] We believe that our corpus-based approach constitutes an important new method for evaluating models of sentence processing. Such models are currently tested exclusively on data obtained for isolated, artificially constructed sentences in controlled lab experiments. The validity of the models can be enhanced considerably if we are able to show that they scale up to model reading data from an eye-tracking corpus, which contains naturally occurring, contextualized sentences. Furthermore, the use of eye-tracking corpora has the advantage of convenience and flexibility: it makes it possible to study arbitrary syntactic constructions, provided that they occur sufficiently frequently in the corpus. There is no need to run a new experiment for every construction or every hypothesis to be investigated.

While the corpus-based approach has great potential, there are limitations as well. The fact that naturally occurring sentences are used means that it is much more difficult to control for confounding factors. In the present paper, we have attempted to include all potentially confounding factors as co-variates in mixed effects models, thus controlling for the influence of these factors. However, it is possible that there are some confounds that we have failed to identify, and therefore they could introduce artifacts in our models. In an experimental setting, the experimenter will often construct materials so that they are matched across conditions, i.e., the sentences only differ in the aspects that the experimenter wants to manipulate, an are identical in all other ways. This reduces the possibility that there are confounding factors that have not been taken into account. Another limitation of the corpus-based approach is data sparseness. No corpus can be so big that it contains all syntactic structures that an experimenter might want to get data on. For example, if we want to investigate prepositional phrase attachment, then there is a good chance that there are enough relevant sentences in the Dundee Corpus. If we want to investigate reduced relative clauses, on the other hand, then probably there are not enough tokens. This situation is even worse if we want to study structures that are ungrammatical or cause serious processing disruption (such as multiple center embeddings). These probably do not occur in the corpus at all. To summarize, experimental data and corpus data have complementary strengths and weaknesses, and should be used in conjunction to maximize the evidence for or against a given theoretical position.

References

Alexopoulou, T., & Keller, F. (2007). Locality, cyclicity and resumption: At the interface between the grammar and the human sentence processor. *Language, 83*(1), 110–160.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, to appear*.

Brysbaert, M., & Vitu, F. (1998). Word skipping: Implications for theories of eye movement control in reading. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 125–147). Oxford: Elsevier.

[8] As an anonymous reviewer points out, there is previous work that uses corpus data to study reading, such as Stevens and Rumelhart's (1975) investigation of reading errors using Augmented Transition Networks. However, these authors did not have eye-tracking corpora at their disposal.

Burnard, L. (1995). *Users guide for the British National Corpus.* British National Corpus Consortium, Oxford University Computing Service.

Clarkson, P. R., & Rosenfeld, R. (1997). Statistical language modeling using the CMU-Cambridge toolkit. In G. Kokkinakis, N. Fakotakis, & E. Dermatas (Eds.), *Proceedings of the 5th European conference on speech communication and technology* (pp. 2707–2710). Rhodes: International Speech Communication Association.

Demberg, V., & Keller, F. (2007). Eye-tracking evidence for integration cost effects in corpus data. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 947–952). Nashville: Cognitive Science Society.

Demberg, V., & Keller, F. (2008). A psycholinguistically motivated version of TAG. In *Proceedings of the 9th international workshop on tree adjoining grammars and related formalisms.* Tübingen.

Elman, J. L. (1991). Distributed reprentations, simple recurrent networks and grammatical structure. *Machine Learning*, *9*, 195–225.

Ferrara 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. *Journal of Eye Movement Research*, *to appear*.

Ferreira, F., & Henderson, J. M. (1993). Reading processes during syntactic analysis and reanalysis. *Canadian Journal of Psychology*, *47*(2), 247–275.

Frisson, S., Rayner, K., & Pickering, M. J. (2006). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 862–877.

Garnsey, S. M., Pearlmutter, N. J., Myers, E. M., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*(1), 58–93.

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68*, 1-76.

Gibson, E. (2000). Dependency locality theory: A distance-dased 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. (2006). The interaction of top-down and bottom-up statistics in the resolution of syntactic category ambiguity. *Journal of Memory and Language*, *54*, 363–388.

Gundel, J., Hedberg, H., & Zacharski. (1993). Referring expressions in discourse. *Language*, *69*, 274–307.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Juhasz, B. J., Starr, M. S., Inhoff, A. W., & Placke, L. (2003). The effects of morphology on the processing of compound words: Evidence from naming, lexical decisions and eye fixations. *British Journal of Psychology*, *94*(2), 223–244.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research*, *45*, 153–168.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*, 580-602.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*(6), 627–645.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lin, D. (1998). An information-theoretic definition of similarity. In J. W. Shavlik (Ed.), *Proceedings of the 15th international conference on machine learning* (pp. 296–304). Madison, WI: Morgan Kaufmann.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 149–157.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

McDonald, S. A., & Shillcock, R. C. (2003a). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*(6), 648-652.

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

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312.

Narayanan, S., & Jurafsky, D. (2002). A Bayesian model predicts human parse preference and reading time in sentence processing. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 59–65). Cambridge, MA: MIT Press.

Padó, U. (2007). *The integration of syntax and semantic plausibility in a wide-coverage model of human sentence processing*. Unpublished doctoral dissertation, Saarland University.

Pickering, M. J., Traxler, M. J., & Crocker, M. W. (2000). Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language*, *43*(3), 447–475.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes*, *41*, 221–250.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, *27*(2), 249–276.

Sampson, G. (1995). *English for the computer: The SUSANNE corpus and analytic scheme*. Oxford: Clarendon Press.

Stevens, A. L., & Rumelhart, D. E. (1975). Errors in reading: An analysis using an augmented transition network model of grammar. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition* (pp. 136–155). San Francisco: W. H. Freeman.

Tabor, W., & Tanenhaus, M. (1999). Dynamical models of sentence processing. *Cognitive Science*, *23*(4), 491–515.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 528–553.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767–794.

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

Witten, I. H., & Bell, T. C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transaction on Information Theory, 37*(4), 1085–1094.

Appendix: Technical Details in Processing the Dundee Corpus

Skipping

The Dundee corpus has a relatively high skipping rate: 45% for first pass reading and 35% for total reading times. This is higher than previously reported numbers, e.g., Brysbaert and Vitu (1998) found a skipping rate of only just over one third in first pass reading. Therefore, many words have a reading time value 0. If we included these data points into our regressions, they would heavily influence the data. This is particularly problematic since the meaning of skipping a word is not the same as the meaning of a very short fixation (closed to 0 ms). Therefore, all the regressions in this article were run on fixated words only, and skipping was dealt with in a separate, logistic regression, which included a binary response variable that specified whether a word was fixated or not. We here only reported the regressions on fixated words because they are more informative.

Track Losses

The rate of track losses is relatively high in the corpus. We define a track loss as a sequence of four adjacent words that are not fixated. Out of the half a million tracked words (approx. 50,000 words $\times$ 10 participants), 7.3% of the data points are invalid due to track losses. We remove them for the regression analyses because the large proportion of track losses otherwise could lead to substantial distortion of the results, in particular for estimating skipping and refixation probabilities.

Spill-Over

Spill-over effects are delays on the target word caused by processing difficulty in the preceding work. We try to capture spill-over effects by including the frequency of the previous word, a flag that indicates whether the previous word was fixated or not, and launch distance as predictors in our models.

Issues Specific to Corpus Data

Newspaper text contains many types of words that are usually not present in specifically designed psycholinguistic experiment items, such as numbers and special characters. We found these words to require special treatment. For example, in our frequency statistics (which we estimated based on word occurrences in the British National Corpus (BNC), after stripping off punctuation), we found an unexpectedly high number of short words with low frequencies (in general we expect that length is negatively correlated with word frequency). We also found that these low frequency words were skipped with higher probability than expected, and received fewer fixations. This indicated that some words were assigned to an inappropriate frequency class. We dealt with this problem by excluding all words that contain numbers, special characters such as punctuation and hyphens, and acronyms (words with more than one capital letter). The variation in the word length of rare words is then considerably lower, and both skipping probability and fixation numbers become monotonous functions, with the rare words skipped least often and fixated (and regressed to) most often.

An alternative treatment of the problematic words would be to change their frequency assignments. For instance, a psycholinguistic reason for changing the frequency of digits would be that they are probably considered as a class of signs in the human processor and therefore should be annotated with their class frequency. Compounds with hyphens on the other hand should not be

annotated with the frequency for the whole compound, as there is evidence in the literature on compound reading that the reading durations of compounds are primarily dependent on the frequency of the first part of the compound (Juhasz, Starr, Inhoff, & Placke, 2003).

Alignment of Tokenizations

Tokenization in the Dundee corpus is often different from the tokenization used by the parsers. Therefore, it is necessary to realign the parsed text with the Dundee corpus segmentation. If a word in the Dundee corpus corresponds to multiple words in the parsed version, we have to combine the theories' predictions for those two words into a single prediction for that token, or split up the Dundee token into two bits. We here decided to combine the predictions for two different words into a single value and use the Dundee corpus tokenization.

For both surprisal and integration cost, we decided to combine predictions by summarization (instead of, e.g., computing the average). Surprisal captures the amount of probability mass of analyses that are not compatible with the current input given the prefix. Two words which are one token in the Dundee corpus (like *we'll*) carry the same information as two separate adjacent tokens (*we* and *'ll*, and thus rule out the same structures, such that the surprisal of *we'll* is exactly the same as the surprisal of *we* plus the surprisal if *'ll* (see Equation (7)).

$$
\begin{aligned}
(7) \quad S_{k+1} + S_{k+2} &= -\log P(w_{k+1}|w_1 \cdots w_k) + -\log P(w_{k+2}|w_1 \cdots w_{k+1}) \\
&= -\frac{\log P(w_1 \cdots w_{k+1})}{P(w_1 \cdots w_k)} - \frac{\log P(w_1 \cdots w_{k+2})}{P(w_1 \cdots w_{k+1})} \\
&= -\log P(w_1 \cdots w_{k+1}) + \log P(w_1 \cdots w_k) - \\
&\quad \log P(w_1 \cdots w_{k+2}) + \log P(w_1 \cdots w_{k+1}) \\
&= \log P(w_1 \cdots w_k) - \log P(w_1 \cdots w_{k+2}) \\
&= -\log \frac{P(w_1 \cdots w_{k+2})}{P(w_1 \cdots w_k)} \\
&= -\log P(w_{k+1}, w_{k+2}|w_1 \cdots w_k) \\
&= S_{k+1,k+2}
\end{aligned}
$$

Similarly, we also decided to add up integration costs, because the relevant quantity is the combined integration cost of the two components, which means that averaging would not be an appropriate measure.

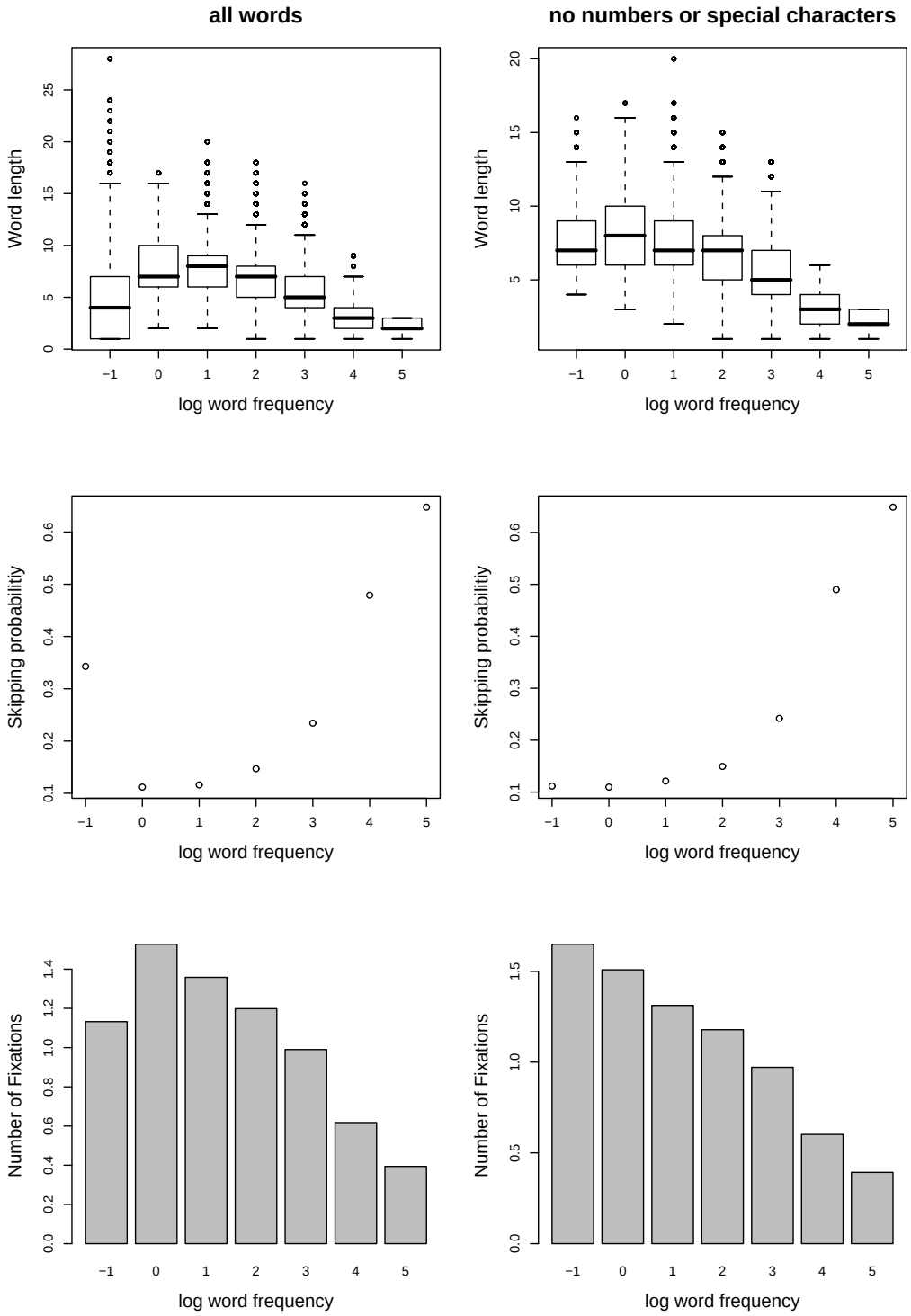


Figure 4. The first column shows word length distributions, skipping probability and numbers of fixation on a word for words of different frequency classes. The second column matches the plots from the first column, but the data set of the second column excludes all words with symbols that are not characters, such as numbers, punctuation, compounds with a hyphen or special signs.