Journal of Eye Movement Research
2(1):1, 1-12

# Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus

Marisa Ferrara Boston
Cornell University

John Hale
Cornell University

Reinhold Kliegl
University of Potsdam

Umesh Patil
University of Potsdam

Shravan Vasishth
University of Potsdam

The *surprisal* of a word on a probabilistic grammar constitutes a promising complexity metric for human sentence comprehension difficulty. Using two different grammar types, surprisal is shown to have an effect on fixation durations and regression probabilities in a sample of German readers' eye movements, the Potsdam Sentence Corpus. A linear mixed-effects model was used to quantify the effect of surprisal while taking into account unigram frequency and bigram frequency (transitional probability), word length, and empirically-derived word predictability; the so-called "early" and "late" measures of processing difficulty both showed an effect of surprisal. Surprisal is also shown to have a small but statistically non-significant effect on empirically-derived predictability itself. This work thus demonstrates the importance of including parsing costs as a predictor of comprehension difficulty in models of reading, and suggests that a simple identification of syntactic parsing costs with early measures and late measures with durations of post-syntactic events may be difficult to uphold.

**Keywords: Reading, eye movements, probabilistic grammar, sentence comprehension, Potsdam Sentence Corpus.**

Reading a sentence involves a succession of fixations and saccades, with information uptake occuring mainly during fixations. The duration of a fixation at a word is known to be affected by a range of word-level factors such as token frequency and empirical predictability as measured in a Cloze task with human subjects (Taylor, 1953; Ehrlich & Rayner, 1981; Kliegl, Grabner, Rolfs, & Engbert, 2004).

When words appear in sentences — as opposed to in isolation — their occurrence is evidently affected by syntactic, semantic and other factors. Research within psycholinguistics over the past half-century has exposed the role of some of these sentence-level factors in accounting for eye movements. Clifton et al. (2007) provides a review of this work, and calls for the development of explicit theories that combine word-level and sentence-level factors. Of course, such combined models would be unnecessary if it turned out that sentence-level factors actually have very little effect on eye movements. These sorts of factors do not figure in current models of eye-movement control such as E-Z Reader (Pollatsek, Reichle, & Rayner, 2006) and SWIFT (Engbert, Nuthmann, Richter, & Kliegl, 2005), whose difficulty predictions derive primarily from statistical properties of individual words and their immediate neighbors.

In this paper, we cast doubt on this simpler view by exhibiting a quantitative model that takes into account both word and sentence-level factors in explaining eye fixation durations and regression probabilities. We show that the surprise value of a word, on a grammar-based parsing model, is an important predictor of processing difficulty independent of factors such as word length, frequency, and empirical predictability. This result harmonizes with the rise of probabilistic theories in psycholinguistics defined over grammatical representations such as constituents and dependency relations (Jurafsky, 1996; Crocker & Brants, 2000; Keller, 2003). In addition to demonstrating the effect of surprisal on eye-movement measures, we also show that surprisal has a small but statistically non-significant effect on empirical predictability.

The paper is organized into three sections. The first section explains the concept of surprisal, summarizing the Hale (2001) formulation. The second section marshals several predictors — surprisal, word length, unigram frequency, bigram frequency (transitional probability in the sense of McDonald & Shillcock, 2003)

and empirical predictability values — in a quantitative model of fixation durations and regression probabilities. We fit this model to the measurements recorded in the Potsdam Sentence Corpus (Kliegl, Nuthmann, & Engbert, 2006), making it possible to determine which predictors account for readers' fixation durations and regressive eye movements. The last section discusses implications of this fitted model for various linking hypotheses between eye movement measures and parsing theories. This final section also discusses the implications of the results for E-Z Reader (Pollatsek et al., 2006) and SWIFT (Engbert et al., 2005).

## Surprisal

Surprisal is a human sentence processing complexity metric; it offers a theoretical reason why a particular word should be easier or more difficult to comprehend at a given point in a sentence. Although various complexity metrics have been proposed over the years (Miller & Chomsky, 1963; Kaplan, 1972; Gibson, 1991; Stabler, 1994; Morrill, 2000; Rohde, 2002; Hale, 2006), surprisal has lately come to prominence within the field of human sentence processing (Park & Brew, 2006; Levy, in press; Demberg & Keller, 2008). This renewal of interest coincides with a growing consensus in that field that both absolute as well as graded grammatical factors should figure in an adequate theory. Surprisal combines both sorts of considerations.

This combination is made possible by the assumption of a probabilistic grammar. Surprisal presupposes that sentence-comprehenders know a grammar describing the structure of the word-sequences they hear. This grammar not only says which words can combine with which other words but also assigns a probability to all well-formed combinations. Such a probabilistic grammar assigns exactly one structure to unambiguous sentences. But even before the final word, one can use the grammar to answer the question: what structures are compatible with the words that have been heard so far? This set of structures may contract more or less radically as a comprehender makes their way through a sentence.

The idea of surprisal is to model processing difficulty as a logarithmic function of the probability mass eliminated by the most recently added word. This number is a measure of the information value of the word just seen as rated by the grammar's probability model; it is nonnegative and unbounded. More formally, define the *prefix probability* of an initial substring to be the total probability of all grammatical[1] analyses that derive $w = w_1 \cdots w_n$ as a left-prefix (definition 1). Where the grammar $G$ and prefix string $w$ (but not $w$'s length, $n$) are understood, this quantity is abbreviated[2] by the forward probability symbol, $\alpha_n$.

$$\text{prefix-probability}(w, G) = \sum_{d \in \mathcal{D}(G, wv)} \text{Prob}(d) = \alpha_n(1)$$

Then the surprisal of the $n^{\text{th}}$ word is the log-ratio of the prefix probability before seeing the word, compared to the prefix probability after seeing it (definition 2).

$$\text{surprisal}(n) = \log_2 \left( \frac{\alpha_{n-1}}{\alpha_n} \right) \tag{2}$$

As the logarithm of a probability, this quantity is measured in bits.

Consider some consequences of this definition. Using a law of logarithms, one could rewrite definition 2 as

$$\log_2(\alpha_{n-1}) - \log_2(\alpha_n)$$

But on a well-defined probabilistic grammar, the prefix probabilities $\alpha$ are always less than one and strictly nonincreasing from left to right. This implies that the two logarithms are to be subtracted in the opposite order. For instance, if a given word brings the prefix probability down from 0.6 to 0.01, the surprise value is 4.09 bits.

Intuitively, surprisal increases when a parser is required to build some low-probability structure. The key insight is that the relevant structure's size need not be fixed in advance as with Markov models. Rather, appropriate probabilistic grammars can provide a larger domain of locality. This paper considers two probabilistic grammars, one based on hierarchical phrase-structure[3] and another based on word-to-word dependencies. These two grammar-types were chosen to il-

[1] In this definition, $G$ is a probabilistic grammar; the only restriction on $G$ is that it provide a set of derivations, $\mathcal{D}$ that assign a probability to particular strings. When $\mathcal{D}(G, u) = \emptyset$ we say that $G$ does not derive the string $u$. The expression $\mathcal{D}(G, wv)$ denotes the set of derivations on $G$ that derive $w$ as the initial part of larger string, the rest of which is $v$. See Jurafsky and Martin (2000), Manning and Schütze (2000) or Charniak (1993) for more details on probabilistic grammars.

[2] Computational linguists typically define a state-dependent forward probability $\alpha_n(q)$ that depends on the particular destination state $q$ at position $n$. These values are indicated in red inside the circles in figure 3(a). It is natural to extend this definition to state sets by summing the state-dependent $\alpha$ values for all members. To define the surprisal of a left-contextualized word on a grammar the summation ranges over all grammatically-licensed parser states at that word's position. The notation $\alpha_n$ (without any parenthesized $q$ argument) denotes this aggregate quantity.

[3] The probabilistic context-free phrase-structure grammars were unlexicalized. See Stolcke (1995) for more information in the methods used in this work. For this purpose, we adapted Levy's implementation of the Stolcke parser, available from http://idiom.ucsd.edu/~rlevy/prefixprobabilityparser.html.

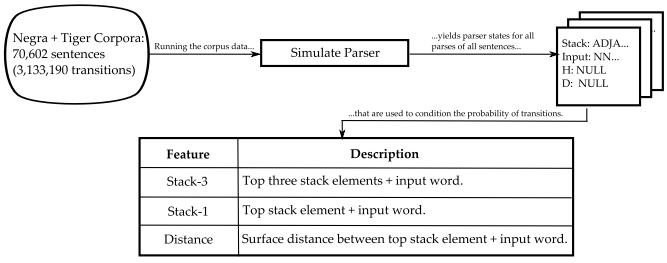
lustrate surprisal's compatibility with different grammar formalisms. Since the phrase-structure approach has already been presented in Hale (2001), the next two sub-sections elaborate the dependency grammar approach.

## *Estimating the parser's probability model*

Consider the German sentence in example 3.

(3) Der alte Kapitaen goss stets ein wenig
the old captain poured always a little
Rum in seinen Tee
rum in his tea
"The old captain always poured a little rum in his tea"

A probabilistic dependency parser can proceed through this sentence from left to right, connecting words that stand in probable head-dependent relationships (Nivre, 2006). In this paper, parser-action probabilities are estimated from the union of two German newspaper corpora, NEGRA (Skut, Krenn, Brants, & Uszkoreit, 1997) and TIGER (König & Lezius, 2003), as in Figure 1.




| Feature | Description |
|---|---|
| Stack-3 | Top three stack elements + input word. |
| Stack-1 | Top stack element + input word. |
| Distance | Surface distance between top stack element + input word. |

*Figure 1.* Estimating the parser's probability model.

Figure 1 defines the method of estimating the parser probabilities from the corpus data. A simulation of the parser is run on the training data, yielding a series of parser states and transitions for all sentences in the corpora. This information informs several features (Hall, 2007), which are then used to condition the probabilities of each transition. A Maximum Entropy training model (Charniak & Johnson, 2005) was used to weight each feature instance for better accuracy.

## *Estimating surprisal*

The prefix probability (definition 1) may be approximated to any degree of accuracy $k$ by summing up the total probability of the top $k$ most probable analyses defined by the dependency parser. Then surprisals can be computed by applying definition 2 following Boston and Hale (2007). Figure 2 shows the surprisals associated with just two of the words in Example 3.

Figure 2 also depicts the dependency relations for this sentence, as annotated in the Potsdam Sentence corpus.[4] Following Tesnière (1959) and Hayes (1964), the word at the arrow head is identified as the 'dependent', the other is the 'head' or 'governor'. The associated part-of-speech tag is written below each actual word; this figures into the surprisal calculation via the parser's probability model. The thermometers indicate surprisal magnitudes; at *alte*, 0.74 bits amounts to very little surprise. In TIGER and NEGRA newspaper text, it is quite typical to see an adjective (ADJA) following an article (ART) unconnected by any dependency relation. By contrast, the preposition *in* is most unexpected. Its surprisal value is 23.83 bits.

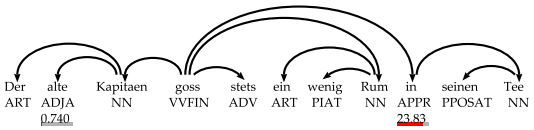



*Figure 2.* Surprisal is a word-by-word complexity metric.

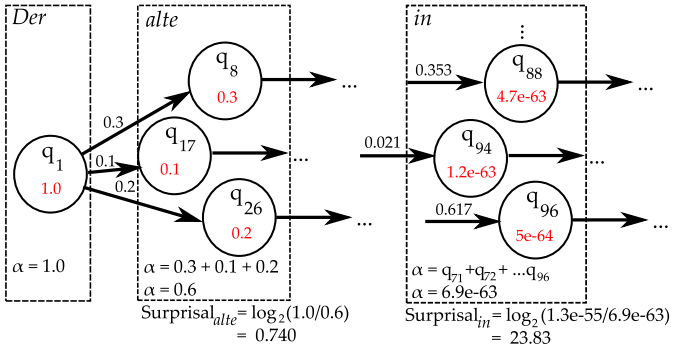
The surprisal values are the result of a calculation that makes crucial reference to instantaneous descriptions of the incremental parser. Figure 3(a) schematically depicts this calculation. At the beginning of Example 3, the parser has seen *der* but the prefix probability is still 1.0 reflecting the overwhelming likelihood that a sentence begins with an article. Hearing the second word *alte*, the top $k = 3$ destination states are, for example, $q_8, q_{17}$ and $q_{26}$ (the state labels are arbitrary). Figure 3(b) reads off the grammatical significance of these alternative destinations: either *alte* becomes a dependent of *der*, or *der* becomes a dependent of *alte* or no dependency predicated. Each transition from state $q_1$ to states $q_8, q_{17}$ and $q_{26}$ has a corpus-estimated probability denoted by the values above the arc (e.g., the transition probability to $q_8$ = 0.3). Approximating definition 1, we find that the total probability of all state trajectories[5] arriving in one of those top 3 is 0.6, and thus the surprisal at *alte* is 0.740 bits.

When the parser arrives at *in*, the prefix probability for the word has made its way down to $6.9 \times 10^{-63}$. Such miniscule probabilities are not uncommon in broad-coverage modeling. What matters for the surprisal calculation is not the absolute value of the prefix probability, but rather the ratio between the old prefix-probability and the new prefix-probability. A high $\alpha_{n-1}/\alpha_n$ ratio means that structural alternatives have been reduced in probability or even completely ruled out since the last word.

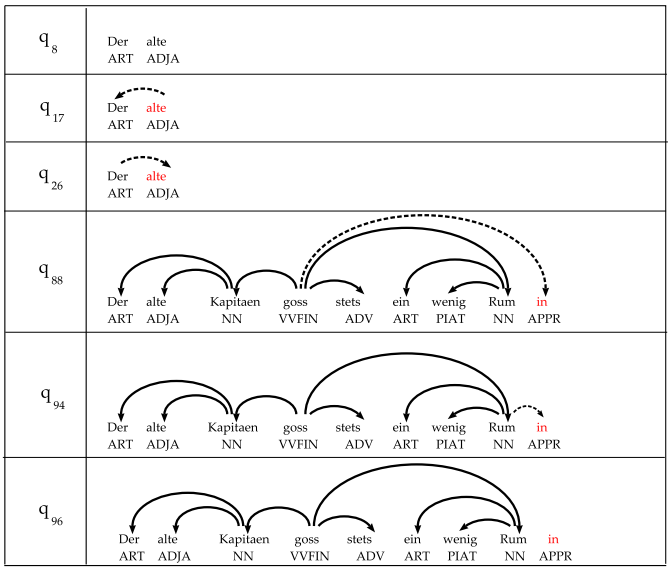
For instance, the action that attaches the preposi-

[4] The labels in the second line (e.g., VVFIN) symbolize the grammatical category for each word as described in the Negra annotation manual (Skut et al., 1997). We are presuming a tagger that accomplishes this task (see Chapter 10 of Manning and Schütze (2000)).

[5] This work takes the Nivre (2006) transition system to be sound and complete with respect to a probabilistic dependency grammar that could, in principle, be written down.




(a) State-based surprisal calculation




(b) Dependency grammar claims in parser states $q$

*Figure 3*. Sketch of surprisal calculation.

tion *in* to its governing verb *goss* is assigned a probability of just over one-third. That action in this left-context leads to the successor state $q_{88}$ with the highest forward probability (indicated inside the circles in red). Metaphorically, the preposition tempers the parser's belief that *goss* has only a single dependent. Of course, *k*-best parsing considers other alternatives, such as state $q_{96}$ in which no attachment is made, in anticipation that some future word will attach *in* as a left-dependent. However these alternative actions are all dominated by the one that sets up the correct *goss*↷*in* dependency. This relationship would be ignored in a 3-gram model because it spans four words. By contrast, this attachment is available to the Nivre (2006) transition system because of its stack-structured memory. In fact, attachments to *stets*, 'always', *ein*, 'a', and *wenig*, 'little', are all excluded from consideration because the parser is projective, i.e., does not have crossing dependencies (Kahane, Nasr, & Rambow, 1998; Buch-Kromann, 2007).

The essence of the explanation is that difficult words force transitions through state-sets whose forward probability is much smaller than at the last word. This explanation is interpretable in light of the linguistic claims made by the parser. However, the explanation is also a numerical one that can be viewed as just another kind of predictor. The next section applies this perspective to modeling observed fixation durations and regression frequencies.

## Predicting eye movements: The role of surprisal

Having sketched a particular formalization of sentence-level syntactic factors in the previous section, this section takes up several other factors (table 1) that figure in models of eye-movement control. Two subsections report answers to two distinct but related questions. The first question is, can surprisal stand in, perhaps only partly, for empirical predictability? If empirical predictability could be approximated by surprisal, this would save eye-movement researchers a great deal of effort; there would no longer be a need to engage in the time-consuming process of gathering predictability scores. Unfortunately, the answer to this first question is negative – including surprisal in a model that already contains word-level factors such as length and bigram frequency does not allow it to do significantly better at predicting empirical predictability scores in the Cloze-type data we considered.

The second question pertains to eye-movement data. The second subsection proceeds by defining a variety of dependent measures commonly used in eye movement research. Then it takes up the question, does adding surprisal as an explanatory factor result in a better statistical model of eye-movement data? The answer here is affirmative for a variety of fixation duration measures as well as regression likelihoods.

### *Does surprisal approximate empirical predictability?*

The Potsdam Sentence Corpus (PSC) consists of 144 German sentences overlayed with a variety of related information (Kliegl, Nuthmann, & Engbert, 2006). One kind of information comes from a predictability study in which native speakers were asked to guess a word given its left-context in the PSC (Kliegl et al., 2004). The probability of correctly guessing the word was estimated from the responses of 272 participants. This diverse pool included high school students, university students, and adults as old as 80 years. As a result of this study, every PSC word — except the first word of each sentence, which has no left context — has associated with it an empirical word-predictability value that ranges from 0 to 1 with a mean (standard deviation) of 0.20 (0.28). These predictability values were submitted to a logit transformation in order to correct for the dependency between mean probabilities

and the associated standard deviations; see (Kliegl et al., 2004) for details.

Table 1 enumerates a set of candidate factors hypothesized to influence logit predictability as sampled in the Kliegl et al. (2004) study. The candidate factors were taken into account simultaneously in a linear mixed-effects model (Pinheiro & Bates, 2000; Bates & Sarkar, 2007; Gelman & Hill, 2007) with sentences as random factors.

The *Deviance Information Criterion* or DIC (Spiegelhalter, Best, Carlin, & Linde, 2002; Spiegelhalter, 2006), (Gelman & Hill, 2007, 524-527) was used to compare the relative quality of fit between models. The DIC depends on the summary measure of fit *deviance* $d = -2 \times$ log-likelihood. Adding a new predictor that represents noise is expected to reduce deviance by 1; more generally, adding $k$ noise predictors will reduce deviance by an amount corresponding to the $\chi^2$ distribution with $k$ degrees of freedom. DIC is the sum of *mean deviance* and $2\times$*the effective number of parameters*; mean deviance is the average of the deviance over all simulated parameter vectors, and the effective number of parameters depends on the amount of pooling in the mixed-effects model. Thus, in mixed-effects models DIC plays the role of the Akaike Information Criterion (Akaike, 1973; Wagenmakers & Farrell, 2004), in which the number of estimated parameters can be determined exactly.

In the linear mixed-effects models, neither version of surprisal showed a statistically significant effect.[6] However, the sign of the coefficient was negative for both variants of surprisal and DIC values were lower when surprisal was added as a predictor. This is as expected: more surprising words are harder to predict. The DIC was 2229 for the simpler model, versus 2220 for each of the two more complex models. Table 2 summarizes the models including surprisal as a predictor.

In sum, the analyses show that surprisal scores exhibit rather weak relations with empirical predictability scores; indeed, they are much weaker than unigram frequency and word length as well as corpus-based bigram frequency. Given the reduction in DIC values, however, including surprisal as part of an explanation for empirical word predictability appears to be motivated. This finding is consistent with the intuition that predictability subsumes syntactic parsing cost, among other factors, although clearly surprisal is not the dominant predictor.

The relation between surprisal and empirical word predictability, though weak, nevertheless raises the possibility that surprisal scores may account for variance in fixation durations independent of the variance accounted for by empirical predictability. We investigate this question next using eye movement data from the Potsdam Sentence Corpus.

## *Does surprisal predict eye movements?*

Surprisal formalizes a notion of parsing cost that appears to be distinct from any similar cost that may be subsumed in empirical predictability protocols. It may thus provide a way to account for eye movement data by bringing in a delimited class of linguistic factors that are not captured by conscious reflection about upcoming words.

To investigate this question empirically, we chose several of the dependent eye movement measures in common use (tables 3 and 4). A distinct class of "first pass" measures reflects the first left-to-right sweep of the eye over the sentence. A second distinction relates to "early" and "late" measures. A widely accepted belief is that the former but not the latter reflect processes that begin when a word is accessed from memory (Clifton et al., 2007, 349). Although these definitions are fairly standard in the literature, controversy remains about the precise cognitive process responsible for a particular dependent measure.

In general, human comprehenders tend to read more slowly under conditions of cognitive duress. For instance, readers make regressive eye movements more often and go more slowly during the disambiguating region of syntactically-ambiguous sentences (Frazier & Rayner, 1982). They also slow down when a phrase must be 'integrated' as the argument of a verb that does not ordinarily take that kind of complement, e.g. "eat justice" provokes a slowdown compared to "eat pizza."

The surprisal complexity metric, if successful in accounting for eye movement data, would fit into the gap between these sorts of heuristic claims and measurable empirical data, alongside computational accounts such as Green and Mitchell (2006), Vasishth and Lewis (2006), Lewis et al. (2006) and Vasishth et al. (in press).

We used the dependent measures in tables 3 and 4 to fit separate linear mixed-effects models that take into account the candidate predictors introduced in the last section: the $n$-gram factors, word length, empirical predictability. For the analysis of regression probabilities (coded as a binary response for each word: 1 signified that a regression occurred at a word, and 0 that it did not occur), we used a generalized linear mixed-effects model with a binomial link function (Bates & Sarkar, 2007), (Gelman & Hill, 2007). Sentences and participants were treated as partially crossed random factors; that is, we estimated the variances associated with differences between participants and differences between sentences, in addition to residual variance of the dependent measures. Then we compared the Deviance Information Criterion value of these simpler models with those of more complex models that had an additional predictor: either surprisal based on the

[6] An absolute t-value of 2 or greater indicates statistical significance at $\alpha = 0.05$. The t-values in a mixed-effects models are only approximations because determining the exact degrees of freedom is non-trivial (Gelman & Hill, 2007).

Table 1
*Candidate explanatory factors for empirical predictability.*

| | |
|---|---|
| | Independent Variables |
| log_freq | logarithm of the token frequency ("unigram") of a word in *Das Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts* (DWDS) (Geyken, 2007; Kliegl, Geyken, Hanneforth, & Würzner, 2006) |
| log_bigram | logarithm of the conditional likelihood of a word given its left neighbor ("bigram") in DWDS |
| length | number of characters in conventional spelling |
| s_dg | surprisal from dependency parser |
| s_cfg | surprisal from phrase-structure parser |
| | Dependent Variable |
| lp | logit-transformed empirical word predictability (Kliegl et al., 2004) |

Table 2
*The effect on logit predictability of log unigram and bigram frequencies, 1/word length, and surprisal computed using the dependency grammar.*

| | Estimate | Std. Error | t-value |
|---|---|---|---|
| | Dependency grammar based surprisal | | |
| (Intercept) | -1.4750 | 0.0312 | -47.3 |
| log_freq | 0.2853 | 0.0353 | 8.1 |
| 1/length | 1.2866 | 0.4028 | 3.2 |
| log_bigram | 0.0671 | 0.0114 | 5.9 |
| s_dg | -0.0009 | 0.0147 | -0.1 |
| | Phrase-structure grammar based surprisal | | |
| (Intercept) | -1.4745 | 0.0312 | -47.3 |
| log_freq | 0.2942 | 0.0358 | 8.2 |
| 1/length | 1.2304 | 0.4047 | 3.0 |
| log_bigram | 0.0639 | 0.0116 | 5.5 |
| s_cfg | -0.0208 | 0.0161 | -1.3 |

Note. An absolute t-value of 2 or greater indicates statistical significance at $\alpha = 0.05$.

Table 3
*Commonly used first-pass dependent measures of eye movement and the stages in parsing processes they are assumed to represent.*

| symbol | measure | definition | hypothesized cognitive process |
|---|---|---|---|
| SFD | single fixation duration | fixation duration on a word during first pass if it is fixated only once | word identification (Clifton et al., 2007, 348) |
| FFD | first fixation duration | time spent on a word, provided that word is fixated during the first pass | word identification |
| FPRT | first-pass reading time or gaze duration | the sum of all fixations in a region during first pass | text integration (Inhoff, 1984) but cf. (Rayner & Pollatsek, 1987) |
| (none) | regression probability | likelihood of jumping back to a previous word during the first pass | resolution of temporary ambiguity (Frazier & Rayner, 1982; Clifton et al., 2003) |

Table 4
*Commonly used non-first-pass dependent measures of eye movement and the stages in parsing processes they are assumed to represent.*

| symbol | measure | definition | hypothesized cognitive process |
|---|---|---|---|
| RPD | regression path duration | the sum of all fixations from the first fixation on the region of interest up to, but excluding, the first fixation downstream from the region of interest | integration difficulty (Clifton et al., 2007, 349) |
| RBRT | right-bounded reading time | summed duration of all fixations in a region of interest, beginning with first pass, including revisits after regressions, and ending before an exit to the right | integration difficulty (Vasishth, Bruessow, Lewis, & Drenhaus, in press) |
| RRT | re-reading time | sum of all fixations after first pass | general comprehension difficulty (Clifton et al., 2007, 363) |
| TRT | total reading time | sum of all fixations | general comprehension difficulty |

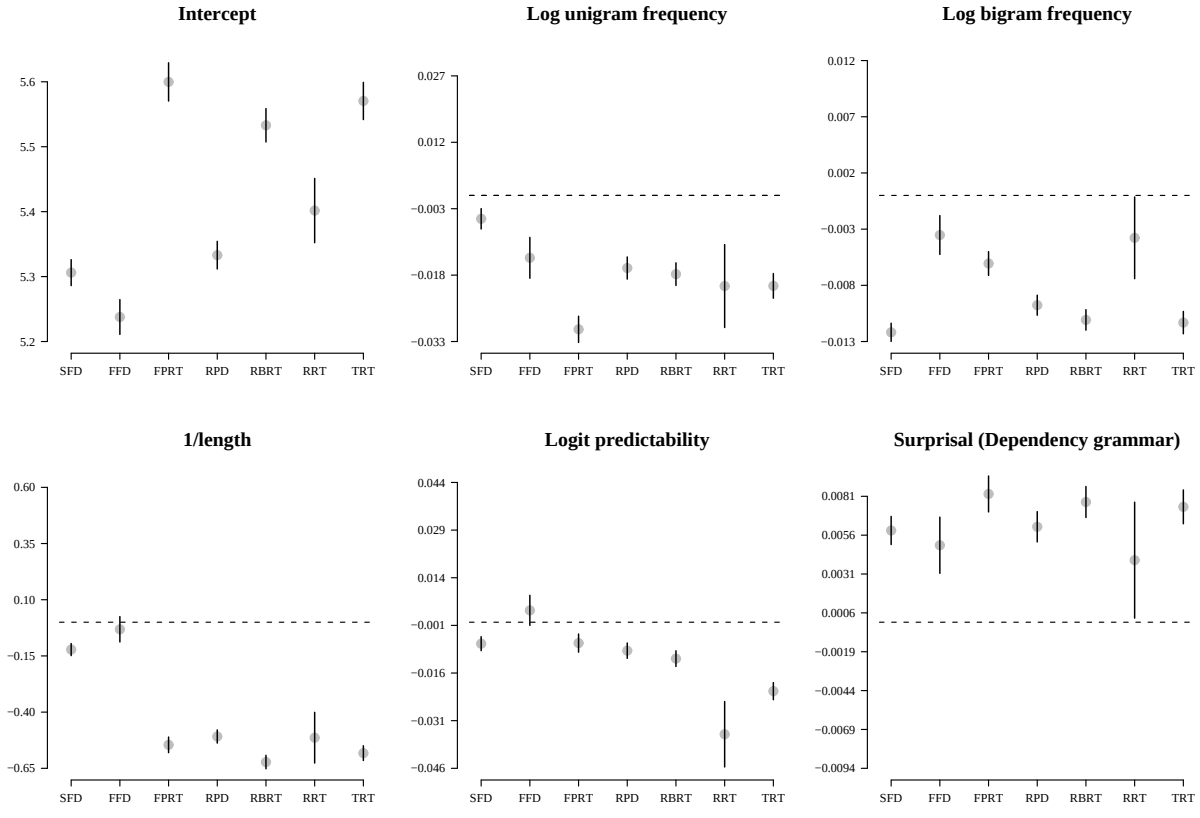



*Figure 4.* Regression coefficients and 95% confidence intervals for the multiple regression using as predictors unigram and bigram frequency, 1/length, logit predictability and dependency grammar based surprisal.

Table 5
*Deviance Information Criterion values for the simpler model, which includes only the word-based statistical measures, and the more complex model, with surprisal added.*

| | unigram+bigram +1/len+pred | +surprisal s.dg | s.cfg |
|---|---|---|---|
| SFD | 43605.9 | 43429.9 | 42966.6 |
| FFD | 19078.4 | 19037.5 | 18945.4 |
| FPRT | 114766.0 | 114600.5 | 114439.0 |
| RPD | 161068.6 | 160860.0 | 160781.5 |
| RBRT | 121072.0 | 120828.3 | 120657.3 |
| RRT | 21916.7 | 21901.6 | 21900.8 |
| TRT | 144511.1 | 144319.3 | 144106.8 |
| Reg | 87028.4 | 87001.7 | 87027.6 |

Table 6
*Log unigram and bigram frequencies, 1/length, and the two surprisal variants as predictors of the so-called early fixation-duration based dependent measures (single-fixation duration and first-fixation duration). All predictors were centered. An absolute t-value of 2 or greater indicates statistical significance at* $\alpha = 0.05$.

| | | Dependency grammar | | | Phrase-structure grammar | | |
|---|---|---|---|---|---|---|---|
| | Predictor | Coef | SE | t-value | Coef | SE | t-value |
| SFD | (Intercept) | 5.3061 | 0.0102 | 519.9 | 5.3099 | 0.0098 | 539.7 |
| | freq | −0.0053 | 0.0012 | −4.5 | −0.0102 | 0.0012 | −8.5 |
| | bigram | −0.0122 | 0.0004 | −29.5 | −0.0106 | 0.0004 | −25.4 |
| | 1/length | −0.1216 | 0.0137 | −8.9 | −0.0815 | 0.0138 | −5.9 |
| | logitpred | −0.0068 | 0.0011 | −6.0 | −0.0051 | 0.0011 | −4.6 |
| | surprisal | 0.0059 | 0.0005 | 12.8 | 0.0133 | 0.0005 | 25.1 |
| FFD | (Intercept) | 5.2378 | 0.0137 | 383.1 | 5.2438 | 0.0123 | 427.3 |
| | freq | −0.0141 | 0.0024 | −6.0 | −0.0165 | 0.0024 | −7.0 |
| | bigram | −0.0035 | 0.0009 | −4.0 | −0.0024 | 0.0009 | −2.8 |
| | 1/length | −0.0315 | 0.0288 | −1.1 | −0.0219 | 0.0288 | −0.8 |
| | logitpred | 0.0037 | 0.0024 | 1.5 | 0.0043 | 0.0024 | 1.8 |
| | surprisal | 0.0050 | 0.0009 | 5.4 | 0.0104 | 0.0009 | 11.0 |

dependency grammar, or surprisal based on phrase-structure grammar.

The calculation of the dependent measures was carried out using the *em* package developed by Logačev and Vasishth (2006). Regarding first-fixation durations, only those values were analyzed that were non-identical to single-fixation durations. In each reading-time analysis reported below, reading times below 50 ms were removed and the dependent measures were log transformed. All predictors were centered in order to render the intercept of the statistical models easier to interpret.

## *Results*

The main results of this paper are summarized in tables 5, 6, 7, and 8. In the multiple regression tables 6-8, a predictor is statistically significant if the absolute t-value is greater than two (p-values are not shown for the reading time dependent measures because in linear mixed-effects models the degrees of freedom are difficult to estimate, Gelman & Hill, 2007).

In order to facilitate comprehension, the multiple regression tables 6-8 are summarized in a more compact form in figures 4 and 5. The graphical summary has the advantage that it is possible, at a glance, to see the consistency in the signs of the coefficients across different measures; the tables will not yield this information without a struggle. The figures are interpreted as follows. The error bars signify 95% confidence intervals for the coefficient estimates; consequently, if an error bar does not cross the zero line, it is statistically significant. This visual test is identical to computing a t-value.

In general, both early and late fixation-duration-based dependent measures exhibited clear effects of unigram frequency, bigram frequency, and logit predictability after statistically controlling for the co-stock of predictors (figures 4, 5). One exception was first-fixation duration (which excludes durations that were also single-fixation durations); here, the effect of pre-

Table 7
*Log unigram and bigram frequencies, 1/length, and the two surprisal variants as predictors of the so-called late measures.*

| | | Dependency grammar | | | Phrase-structure grammar | | |
|---|---|---|---|---|---|---|---|
| | Predictor | Coef | SE | t-value | Coef | SE | t-value |
| RPD | (Intercept) | 5.5998 | 0.0150 | 373.0 | 5.6268 | 0.0146 | 386.2 |
| | freq | −0.0302 | 0.0015 | −20.0 | −0.0328 | 0.0015 | −21.4 |
| | bigram | −0.0061 | 0.0005 | −11.2 | −0.0051 | 0.0005 | −9.4 |
| | 1/length | −0.5456 | 0.0177 | −30.8 | −0.5173 | 0.0178 | −29.0 |
| | logitpred | −0.0066 | 0.0015 | −4.5 | −0.0053 | 0.0015 | −3.6 |
| | surprisal | 0.0083 | 0.0006 | 14.0 | 0.0110 | 0.0007 | 16.6 |
| RBRT | (Intercept) | 5.5330 | 0.0131 | 421.7 | 5.5549 | 0.0127 | 435.8 |
| | freq | −0.0178 | 0.0013 | −13.6 | −0.0208 | 0.0013 | −15.7 |
| | bigram | −0.0111 | 0.0005 | −23.8 | −0.0100 | 0.0005 | −21.2 |
| | 1/length | −0.6223 | 0.0153 | −40.7 | −0.5921 | 0.0154 | −38.5 |
| | logitpred | −0.0115 | 0.0013 | −9.1 | −0.0102 | 0.0013 | −8.1 |
| | surprisal | 0.0077 | 0.0005 | 15.2 | 0.0115 | 0.0006 | 20.1 |
| RRT | (Intercept) | 5.4017 | 0.0253 | 213.8 | 5.4451 | 0.0222 | 245.3 |
| | freq | −0.0205 | 0.0048 | −4.3 | −0.0163 | 0.0048 | −3.4 |
| | bigram | −0.0038 | 0.0019 | −2.0 | −0.0052 | 0.0019 | −2.8 |
| | 1/length | −0.5137 | 0.0576 | −8.9 | −0.5280 | 0.0577 | −9.2 |
| | logitpred | −0.0353 | 0.0053 | −6.7 | −0.0354 | 0.0053 | −6.7 |
| | surprisal | 0.0040 | 0.0019 | 2.1 | −0.0048 | 0.0021 | −2.3 |
| TRT | (Intercept) | 5.5705 | 0.0146 | 381.1 | 5.5877 | 0.0142 | 392.8 |
| | freq | −0.0204 | 0.0014 | −14.4 | −0.0241 | 0.0014 | −16.7 |
| | bigram | −0.0113 | 0.0005 | −22.3 | −0.0101 | 0.0005 | −19.6 |
| | 1/length | −0.5825 | 0.0167 | −34.9 | −0.5496 | 0.0168 | −32.8 |
| | logitpred | −0.0217 | 0.0014 | −15.8 | −0.0203 | 0.0014 | −14.8 |
| | surprisal | 0.0074 | 0.0006 | 13.4 | 0.0124 | 0.0006 | 19.8 |

Table 8
*Log unigram and bigram frequencies, 1/length, and the two surprisal variants as predictors of regression probabilities.*

| | Dependency grammar | | | | Phrase-structure grammar | | | |
|---|---|---|---|---|---|---|---|---|
| Predictor | Coef | SE | z-score | p-value | Coef | SE | z-score | p-value |
| (Intercept) | −2.4117 | 0.0801 | −30.1 | <0.01 | −2.2530 | 0.0744 | −30.3 | <0.01 |
| log_freq | −0.2133 | 0.0114 | −18.7 | <0.01 | −0.2076 | 0.0116 | −17.9 | <0.01 |
| bigram | 0.0880 | 0.0042 | 21.2 | <0.01 | 0.0859 | 0.0042 | 20.5 | <0.01 |
| len | 0.2916 | 0.1317 | 2.2 | 0.03 | 0.3043 | 0.1320 | 2.3 | 0.02 |
| logitpred | 0.0422 | 0.0110 | 3.8 | <0.01 | 0.0442 | 0.0110 | 4.0 | <0.01 |
| surprisal | 0.0236 | 0.0045 | 5.2 | <0.01 | 0.0045 | 0.0050 | 0.9 | 0.37 |

dictability and the reciprocal of length was not significant.

These simpler models were augmented with one of two surprisal factors, one based on dependency grammar, the other based on phrase-structure grammar. As summarized in the table 5, for virtually every dependent measure the predictive error (DIC value) was lower in the more complex model that included surprisal. One exception was regression probability, in which the phrase-structure based grammar predictions did not reduce DIC.

For fixation durations (tables 6, 7 and figures 4, 5), in general both versions of surprisal had a significant effect in the predicted direction (that is, longer durations for higher surprisal values). One exception was the effect of phrase-structure based surprisal on rereading time; here, reading time was longer for lower surprisal values. However, since the rereading time data is sparse (about 1/10th of the other measures; the sparseness of the data is also reflected in the relatively wide confidence intervals for the coefficient estimates of rereading time), it may be difficult to interpret this result, especially given the consistently positive coefficients for surprisal in all other dependent measures.

For regression probabilities (table 8), dependency-grammar based surprisal had a significant effect over

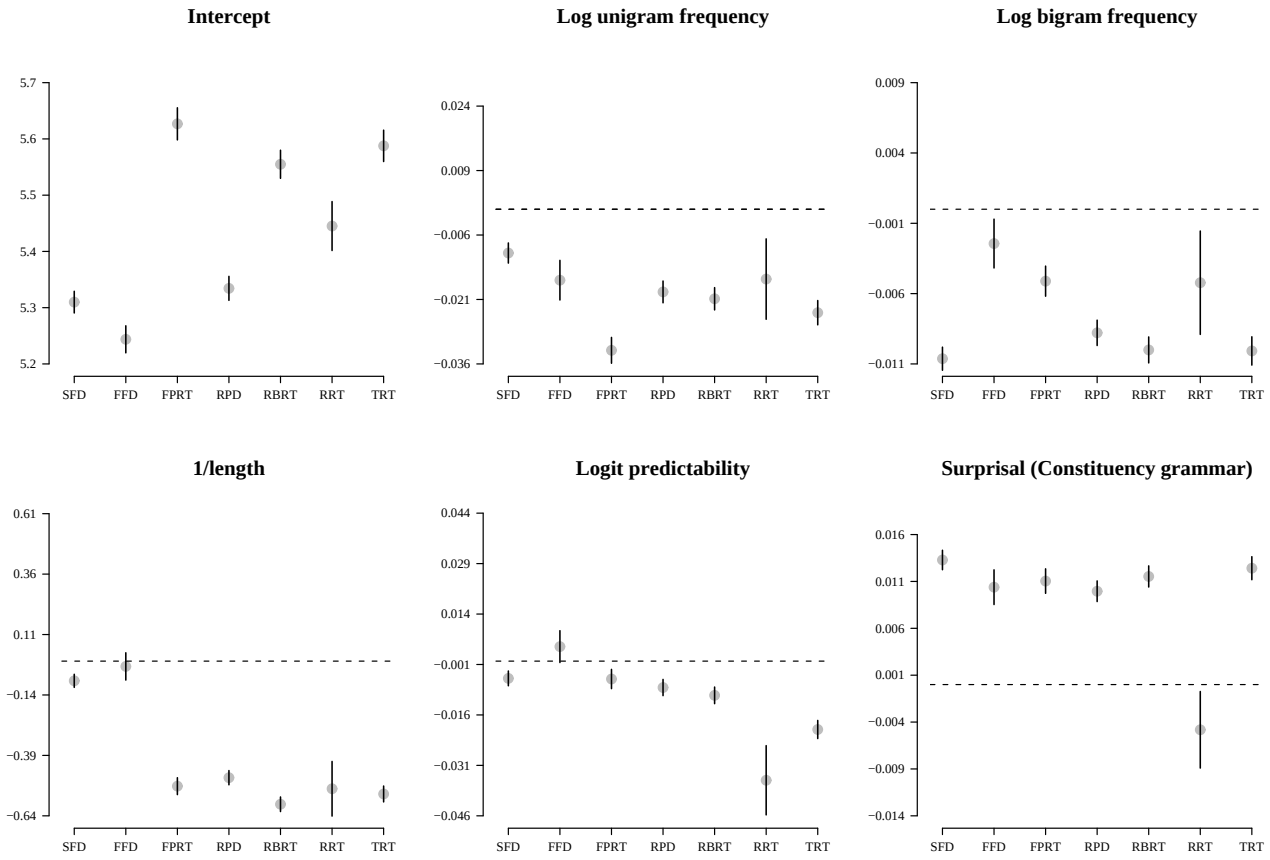



*Figure 5*. Regression coefficients and 95% confidence intervals for the multiple regression, using as predictors unigram and bigram frequency, 1/length, logit predictability and phrase-structure based surprisal.

and above the other predictors: an increase in surprisal predicts a greater likelihood of a regression. Phrase-structure based surprisal is not a significant predictor of regression probability, but the sign of the coefficient is also negative, as in the dependency-based model.

## Discussion

The work presented in this paper showed that surprisal values calculated with a dependency grammar as well as with a phrase-structure grammar are significant predictors of reading times and regressions. The role of these surprisals as predictors was still significant even when empirical word predictability, *n*-gram frequency and word length were also taken into account. On the other hand, surprisal did not appear to have a significant effect on empirical predictability as computed in eye-movement research.

The high-level factor, surprisal, appears in both the so-called early and late measures, with comparable magnitudes of the coefficients for surprisal. This finding is thus hard to reconcile with a simple identification of early measures with syntactic parsing costs and late measures with durations of post-syntactic events. It may be that late measures include the time-costs of syntactic processes initiated much earlier.

The early effects of parsing costs are of high relevance for the further development of eye-movement control models such as E-Z Reader (Pollatsek et al., 2006) and SWIFT (Engbert et al., 2005). In these models, fixation durations at a word are a function of word-identification difficulty, which in turn is assumed to be dependent on word-level variables such as frequency, length and predictability. Although these variables can account for a large proportion of the variance in fixation durations and other measures, we have shown that surprisal plays an important role as well. Of these three predictors, empirical predictability is an "expensive" input variable because it needs to be determined in an independent norming study and applies only to the sentences used in this study. This fact greatly limits the simulation of eye movements collected on new

sentences. It had been our hope that surprisal measures (which can also be computed from available treebanks) could be used as a generally available substitute of empirical predictability. Our results did not match these expectations for the two types of surprisal scores examined here. Nevertheless, given the computational availability of surprisal values, it is clearly a candidate for being included as a fourth input variable in future versions of computational models. As Clifton et al. (2007) note, no model of eye-movement control currently takes factors such as syntactic parsing cost and semantic processing difficulty into account. While some of this variance is probably captured indirectly by empirical predictability, the contribution of this paper is to demonstrate how syntactic parsing costs can be estimated using probabilistic knowledge of grammar.

## References


Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *2nd international symposium on information theory* (pp. 267–281). Budapest, Hungary.

Bates, D., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes (R package version 0.9975-11) [Computer software].

Boston, M. F., & Hale, J. T. (2007). Garden-pathing in a statistical dependency parser. In *Proceedings of the Midwest Computational Linguistics Colloquium.* West Lafayette, IN: Midwest Computational Linguistics Colloquium.

Buch-Kromann, M. (2007). Dependency-based machine translation and parallel parsing without the projectivity and edge-factoring assumptions. a white paper. In *Copenhagen business school working papers.* Denmark: Copenhagen Business School. CBS. (ISBN x656555814)

Charniak, E. (1993). *Statistical language learning*. Cambridge, MA: MIT Press.

Charniak, E., & Johnson, M. (2005, June). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 173–180). Ann Arbor, Michigan: Association for Computational Linguistics.

Clifton, C., Juhasz, B., Ashby, J., Traxler, M. J., Mohamed, M. T., Williams, R. S., et al. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory and Language*, *49*, 317–334.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye Movements in Reading Words and Sentences. In R. V. Gompel, M. Fisher, W. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 341–372). Amsterdam: Elsevier.

Crocker, M. W., & Brants, T. (2000). Wide-coverage probabilistic sentence processing. *Journal of Psycholinguistic Research*, *29(6)*, 647–669.

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

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*, 641–655.

Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*, 777–813.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Ed.), *Idioms and Collocations: Corpus-based Linguistic, Lexicographic Studies.* London: Continuum Press.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Unpublished doctoral dissertation, Carnegie Mellon University.

Green, M., & Mitchell, D. (2006). Absence of real evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language*, *55*(1), 1–17.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 1–8). Pittsburgh, PA: Carnegie Mellon University.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 609–642.

Hall, K. (2007, June). K-best spanning tree parsing. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 392–399). Prague, Czech Republic: Association for Computational Linguistics.

Hayes, D. (1964). Dependency theory: A formalism and some observations. *Language*, *40*, 511–525.

Inhoff, A. W. (1984). Two stages of word processing during eye fixations in the reading of prose. *Journal of verbal learning and verbal behavior*, *23*(5), 612–624.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognition*, *20*, 137–194.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kahane, S., Nasr, A., & Rambow, O. (1998). Pseudo-projectivity: A polynomially parsable non-projective dependency grammar. In *Proceedings of COLING-ACL* (pp. 646–652). Montréal, Canada: Université de Montréal.

Kaplan, R. M. (1972). Augmented transition networks as psychological models of sentence comprehension. *Artificial Intelligence*, *3*, 77–100.

Keller, F. (2003). A probabilistic parser as a model of global processing difficulty. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th annual conference of the cognitive science society* (pp. 646–651). Boston, MA.

Kliegl, R., Geyken, A., Hanneforth, T., & Würzner, K. (2006). *Corpus matters: A comparison of German DWDS and CELEX lexical and sublexical frequency norms for the prediction of reading fixations.* Unpublished manuscript.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, *16*, 262–284.

Kliegl, R., Nuthmann, A., & Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General*, *135*, 12–35.

König, E., & Lezius, W. (2003). *The TIGER language - a description language for syntax graphs, Formal definition* (Tech. Rep.). Germany: IMS, Universität Stuttgart.

Levy, R. (in press). Expectation-based syntactic comprehension. *Cognition*.

Lewis, R. L., Vasishth, S., & Van Dyke, J. (2006, October). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science*, *10*(10), 447–454.

Logačev, P., & Vasishth, S. (2006). *The em package for computing eyetracking measures* (Tech. Rep.). Germany: Universität Potsdam.

Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

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

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 419–491). NY: John Wiley.

Morrill, G. (2000). Incremental processing and acceptability. *Computational Linguistics*, *26*(3), 319–338.

Nivre, J. (2006). *Inductive dependency parsing*. Dordrecht, Netherlands: Springer.

Park, J., & Brew, C. (2006). A finite-state model of human sentence processing. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL* (pp. 49–56). Sydney, Australia.

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

Pollatsek, A., Reichle, E., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, *52*, 1–56.

Rayner, K., & Pollatsek, A. (1987). Eye movements in reading: A tutorial review. *Attention and performance XII: The psychology of reading*, 327–362.

Rohde, D. L. (2002). *A connectionist model of sentence comprehension and production*. Unpublished doctoral dissertation, Carnegie Mellon University.

Skut, W., Krenn, B., Brants, T., & Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*. Washington, DC.

Spiegelhalter, D. J. (2006). Two brief topics on modelling with WinBUGS. In *Icebugs conference proceedings*. Hanko, Finland.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. van der. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, *64*(B), 583–639.

Stabler, E. (1994). The finite connectivity of linguistic structure. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing* (pp. 303–336). Hillsdale, NJ: Lawrence Erlbaum.

Stolcke. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, *21*, 165–202.

Taylor, W. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly*, *30*, 415–433.

Tesnière, L. (1959). *Eléments de syntaxe structurale*. Paris: Editions Klincksiek.

Vasishth, S., Bruessow, S., Lewis, R. L., & Drenhaus, H. (in press). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*.

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

Wagenmakers, E., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review*, *11*(1), 192–196.