Surprisal-based comparison between a symbolic and a connectionist model of sentence processing

Stefan L. Frank (S.L.Frank@uva.nl)
Institute for Logic, Language and Computation
University of Amsterdam
Science Park 904, 1098 XH Amsterdam, The Netherlands

Abstract

The 'unlexicalized surprisal' of a word in sentence context is defined as the negative logarithm of the probability of the word's part-of-speech given the sequence of previous parts-of-speech of the sentence. Unlexicalized surprisal is known to correlate with word reading time. Here, it is shown that this correlation grows stronger when surprisal values are estimated by a more accurate language model, indicating that readers make use of an objectively accurate probabilistic language model. Also, surprisals as estimated by a Simple Recurrent Network (SRN) were found to correlate more strongly with reading-time data than surprisals estimated by a Probabilistic Context-Free Grammar (PCFG). This suggests that the SRN forms a more accurate psycholinguistic model.

Keywords: Surprisal theory; Sentence processing; Reading time; Probabilistic Context-Free Grammar; Simple Recurrent Network.

Introduction

The time needed to read a word in sentence context depends on many of the word's properties, ranging from low-level features such as the word's frequency, to high-level properties such as its effect on the syntactic and semantic representation of the sentence as a whole. Based on earlier work by Hale (2001), Levy (2008) suggested that many of these aspects can be combined into a single 'causal bottleneck': the word's *surprisal*. Formally, the surprisal of word w_t occurring at position t in a sentence is

$$\text{surprisal}(w_t) = -\log(\Pr(w_t|w_{1\ldots t-1})), \qquad (1)$$

where $\Pr(w_t|w_{1\ldots t-1})$ is the probability of w_t given the sentence's previous words $w_{1\ldots t-1}$. Informally, a word's surprisal can be viewed as the extent to which its occurrence came unexpected. According to so-called surprisal theory, there is a positive linear relation between the time needed to read w_t and its surprisal: Less expected words take longer to read.

Surprisal theory can be derived from particular assumptions about sentence processing in at least two different ways. Levy (2008) showed that it follows from the assumptions that a sentence-so-far is mentally represented as a probability distribution over all interpretations of complete sentences, and that word reading time reflects the extent to which this distribution needs to be updated. Alternatively, Smith and Levy (2008) argue that readers attempt to minimize expected processing cost. Under certain additional assumptions, it follows that word reading time is linearly related to word surprisal, irrespective of how readers develop their expectations about upcoming words.

It is of particular interest to the current paper that neither of these derivations of surprisal theory depends on the nature of the probability model that is used to estimate the surprisal values. Levy (2008) uses a Probabilistic Context-Free Grammar (PCFG) whereas Smith and Levy (2008) take a trigram model, but both these choices seem based on practical rather than theoretical considerations. In the work presented here, surprisal values are estimated by a PCFG and by a Simple Recurrent Network (SRN; Elman, 1990).

In spite of the theoretical arguments mentioned above, empirical evidence for surprisal theory is still quite scarce. Hale (2001) and Levy (2008) give examples of how the theory accounts for particular psycholinguistic phenomena, but they do not perform any comparison between surprisal values and reading-time data on a scale that allows for investigating whether there is a statistically significant relation in general.

In contrast, Smith and Levy (2008) look at reading-time measurements over a collection of British newspaper articles and show that there is indeed an (approximately) linear relation between surprisals and reading times. However, they did not investigate whether this effect exceeds that of the words' 'forward transitional probabilities' $\Pr(w_t|w_{t-1})$, which correlate with surprisal.

Taking these transitional probabilities into account, Demberg and Keller (2008) did not find a statistically significant positive effect (and in some cases even a significant negative effect) of surprisal on reading times on English newspaper texts. They did, however, discover a weak but positive relation between reading times and what they called 'unlexicalized' or 'structural surprisal', which is the surprisal of the words' part-of-speech, given the parts-of-speech of the previous words in the sentence (i.e., the *w*s in Equation 1 represent parts-of-speech rather than words). Boston, Hale, Patil, Kliegl, and Vasishth (2008) replicated this finding using a German data set.

All in all, there is little evidence that word surprisal explains reading-time data in general, above what is already explained by forward transitional probability. *Unlexicalized* surprisal, on the other hand, does seem to correlate positively with reading times.

Before accepting this conclusion, however, we need to be careful not to conflate two different interpretations of the concept 'surprisal'. First, there is an objective sense of surprisal, according to which it is a value that can be observed, or at least reliably estimated. In this sense, $\Pr(w_t|w_{1\ldots t-1})$ of Equation 1 is estimated using some probability model that is be-

lieved to be sufficiently accurate. All the authors mentioned above subscribe to such an interpretation when they estimate surprisal by training some well-defined probability model on text corpora. Second, there is a subjective sense of surprisal, according to which it is an unobservable psychological variable. In this sense, $\Pr(w_t|w_{1\ldots t-1})$ expresses the extent to which the *reader* expects word w_t. Since word processing depends on the reader's expectations, it is this subjective interpretation that matters when surprisal is claimed to affect reading.

It is not at all certain that the reader's expectations follow those of the probability models used to estimate surprisal values. Yet, current research implicitly assumes that a reader's expectations about upcoming words (or parts-of-speech) can be reliably estimated from text corpora. In other words, the two senses of surprisal are taken to be one and the same.

The first aim of this paper is to investigate whether this assumption is warranted. Indeed, it will be shown that more accurate probability models also predict reading times more accurately, thereby providing further support for (unlexicalized) surprisal theory.

Assuming that surprisal theory holds, the paper's second goal is to use the theory for investigating what type of probability model best describes the human sentence-processing system. The psychological validity of two simple sentence-processing models (one symbolic, the other connectionist) is evaluated by comparing their surprisal estimates to reading-time data. As it turns out, the connectionist model provides a more accurate account of the data, suggesting that it forms a better description of human sentence processing than does the symbolic model.

Language models and psycholinguistic models

By definition, a probabilistic language model provides an estimate of the probability $\Pr(w_{1\ldots t})$ of any sentence-initial word sequence $w_{1\ldots t}$. Turning these probabilities into surprisal values is trivial, since

$$-\log(\Pr(w_t|w_{1\ldots t-1})) = \log(\Pr(w_{1\ldots t-1})) - \log(\Pr(w_{1\ldots t})).$$

The accuracy of a language model can be measured by having it generate surprisal values for each word in some validation text. If the model is accurate, it will assign a high probability (i.e., low surprisal) to the text's words. Therefore, the average surprisal over all these words can be taken as a measure for language model accuracy. The lower the average surprisal, the more accurate the language model.[1]

An accurate language model is not necessarily an accurate psycholinguistic model. As explained in the Introduction, this is because the extent to which words are surprising to a reader may deviate from the surprisal values as estimated by the language model. A language model forms an accurate psycholinguistic model if its surprisal estimates account for observed reading times. The coefficient of correlation between surprisals and reading times can therefore be taken as a measure of psycholinguistic model accuracy.

In short, the accuracy of a model that generates surprisal estimates can be evaluated in two different ways: It is an accurate language model if the average surprisal estimate is low, and it is an accurate psycholinguistic model if the surprisal estimates correlate strongly (and positively) with reading times. Only if, in general, more accurate language models are also more accurate psycholinguistic models, can we conclude that human readers use something like an objectively accurate language model for sentence processing.

Method

Two basic models of sentence processing will be investigated: a PCFG (a standard symbolic model) and a SRN (the quintessential connectionist model). Both are trained and tested on newspaper texts (or, rather, on the part-of-speech (pos-)tags from these texts). Each model will estimate not just one, but a whole range of surprisal values for each pos-tag of the test texts. This means that there is also a range of model accuracies, which allows for investigating the relation between langauge model accuracy and psycholinguistic model accuracy.

Text corpora

Training corpus Both models were trained on the *Wall Street Journal* (WSJ) part of the Penn Treebank, which is the same corpus as used by Demberg and Keller (2008). It comprises 49,208 sentences annotated with their syntactical tree structure. Since only unlexicalized surprisal is investigated here, all words were removed, that is, pos-tags were used as lexical items. There are 45 different parts-of-speech, including punctuation marks, parentheses, etcetera.

Test corpus The models were tested on the Dundee corpus (Kennedy & Pynte, 2005), which was also used by Demberg and Keller (2008) and Smith and Levy (2008). It consists of 2,368 sentences that appeared in *The Independent* newspaper. Pos-tags for these sentence were automatically generated by the pos-tagger developed by Brill (1993). These tags were checked by hand, and where necessary adapted according to the Penn Treebank pos-tagging guidelines (Santorini, 1991).

The Dundee corpus comes with eye-tracking data from ten subjects. Different reading-time measures can be extracted from these data. Here, we use only first-pass reading time, defined as the total fixation time on a word before any fixation on a later word of the same sentence.

The models

Probabilistic Context-Free Grammar A PCFG gives rise to surprisal values because it assigns a probability to each sentence-initial word string $w_{1\ldots t}$, which, as mentioned above, can be transformed into surprisals. The probability of $w_{1\ldots t}$ is the sum of probabilities of all sentences that start with $w_{1\ldots t}$. The probability of a sentence is the total probability of all its

[1] Incidentally, the average surprisal is an estimate of the language entropy.

possible tree structures, and the probability of a tree equals the product of probabilities of all the production rules involved in its construction.

Following Demberg and Keller (2008), the PCFG's rules and their probabilities were induced from the WSJ treebank using an algorithm developed by Roark (2001). Next, Roark's top-down incremental parser was applied to the pos-tag sequences from the Dundee corpus. At each point of a sentence, the parser comes up with not just one but many (partial) parse trees, each with an associated probability. To reduce computational load, many of these trees are discarded. Simply stated, if the current most likely parse has a probability p, then only partial parses with a probability larger than $10^{-\theta}p$ are kept. Parameter θ controls the so-called 'beam width'. The larger the beam width, the more parses are taken into account, so the more accurate (and slower) the parser becomes.

The beam width affects not only the generated parses and their probabilities, but also the estimated surprisal of each sentence's pos-tags. A range of surprisal estimates can therefore be obtained by varying the beam width. Here, θ was varied between 1 and 16.

Simple Recurrent Network SRNs are commonly used for next-word prediction. In this task, the network is given an input sentence, one word at a time, and has to predict at each point which word will be the next input. Usually, there is one output unit for each word type, and output activations are forced to sum to 1 so that they can be interpreted as probabilities: After processing the input sequence $w_{1\ldots t-1}$, the activation of the output unit representing word w_t is the network's estimate of $\Pr(w_t|w_{1\ldots t-1})$. Taking the negative logarithm of this activation yields the estimated surprisal of w_t.

An SRN was trained to predict each next pos-tag in the pos-tag sequences from the WSJ corpus (ignoring the tree structures).[2] The network had 45 input and output units (one for each part-of-speech type) and 100 hidden units. For a fair comparison with the PCFG model, two small adaptations to the standard SRN were introduced. First, hidden-unit activations were reset at the beginning of each sentence. This makes sure that each sentence is independent from all others, as is the case for the PCFG model. Second, the network was trained to also estimate a probability for each sentence's first part-of-speech, as does the PCFG.

To obtain a range of surprisal values, the network was tested on pos-tag sequences from the Dundee corpus at several moments during training: after every 1,000 training sentence until sentence number 5,000; from thereon after every 5,000 sentences; and finally after training on all 49,208 WSJ sentences.

Statistical notes

Nearly all methods for statistical analysis assume independence among observations. Although complete independence may be impossible to obtain in practice, this need not to be too problematic as long as experiments can be set up in a way that dependencies among observations are minimized. The current case, however, is fundamentally different: If independence among words in a sentence is assumed, the concept of surprisal becomes meaningless because a word's surprisal depends on the sentence's previous words. Hence, a statistical analysis of word (or pos-tag) surprisal relies on an independence assumption that is inconsistent with the concept of surprisal itself.

It is difficult to estimate how serious this problem is in practice. To make sure that we can rely on the outcome of the analyses, the unit of observation was changed from words to sentences. This does not mean that sentences are truly independent from one another. Rather, both the PCFG and the SRN *treat* sentences as independent. By taking sentence reading times as data points, the independence assumption of the statistical analysis becomes consistent with the models' assumption of independence between sentences. Taking reading times at the sentence rather than word level has the additional advantage of doing away with possible spill-over effects (i.e., surprisal effects that show up with some delay) taking place within a sentence.

The reading time for a sentence (from here on denoted as RT) is the sum of first-pass reading times for all words in the sentence, divided by the sentence's number of characters to compensate for differences in sentence length. Likewise, the surprisal of a sentence is defined as the average surprisal of its pos-tags. If the sentence consists of n pos-tags $w_{1\ldots n}$:

$$\begin{aligned}\text{surprisal}(w_{1\ldots n}) &= -\frac{1}{n}\sum_{t=1}^{n}\log(\Pr(w_t|w_{1\ldots t-1})) \qquad (2)\\ &= -\frac{1}{n}\log(\Pr(w_{1\ldots n})).\end{aligned}$$

Data points (i.e., sentence/subject-combinations) were removed from the analysis if:

- The sentence could not be parsed by the PCFG at all levels of θ;
- The subject fixated on fewer than four words (this includes all sentences with fewer than four words);
- There was a track loss, defined as more than three consecutive non-fixated words;
- After removing data points for the reasons above, log(RT) was more than three standard deviations from the mean for that subject.

This left a total of 16,755 data points (between 1,330 and 1,825 per subject).

Results

Language model accuracy

The *in*accuracy of the language model is the average sentence surprisal (Equation 2), weighted by the number of times the

[2] Training sequences were presented in random order. Output units had softmax activation functions and cross-entropy error was minimized by the standard backpropagation algorithm.

sentence takes part in the analysis. This number varies over sentences, for example because track losses occurred for particular sentence-subject combinations.

Figure 1 shows the range of language model accuracies for the PCFG and SRN models. As expected, the PCFG model becomes more accurate as beam width increases. Likewise, the SRN model becomes more accurate over the course of training.

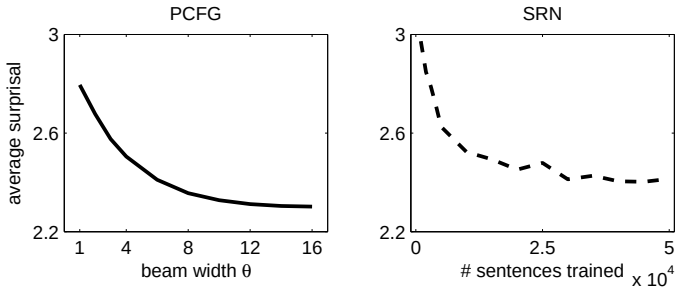

Figure 1: Average surprisal (i.e., inaccuracy of language model). Left: For PCFG, as a function of beam-width θ. Right: For SRN, as a function of the number of sentences trained on.

Psycholinguistic model accuracy

First, RTs were centered by subtracting the average RT for each subject from that subject's RTs. This does away with individual differences in general reading speed. Next, a single measure for psycholinguistic model accuracy was obtained for all subjects by computing the coefficient of correlation between these centered RTs and surprisals values.

Figure 2 shows how the accuracy of the language model is related to the accuracy of the psycholinguistic model. First, as predicted by unlexicalized surprisal theory, there is a significant positive correlation between sentence surprisal and RT, at least for the more accurate languages models. Second, the relation between language model accuracy and psycholinguistic model accuracy is nearly monotonous: Lower average surprisal results in a stronger correlation between surprisal and RT. In other words, accurate language models are generally also accurate psycholinguistic models. This is in line with the assumption that readers use an objectively accurate language model.

Comparing PCFG and SRN

As can be seen in Figure 2, the SRN accounts for more of the variance in RT than does the PCFG, even where the two models have the same language model accuracy. When comparing the SRN at the end of training to the best PCFG (i.e., with $\theta = 16$), the correlation between SRN-based surprisals and RTs is significantly stronger than the correlation between PCFG-based surprisals and RTs ($t = 3.52; p < .001$).[3]

[3]The correlation between the two sets of surprisal estimates is .85.

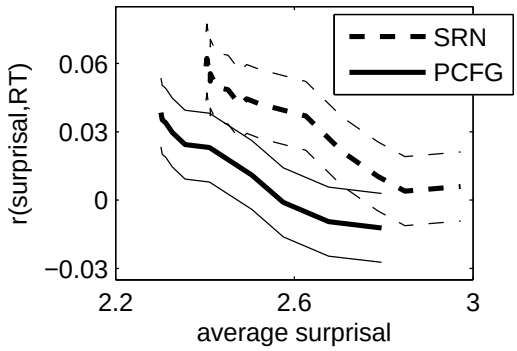

Figure 2: Correlation between surprisal and RT (i.e., accuracy of psycholinguistic model) for PCFG and SRN as a function of language model inaccuracy. Thin lines indicate 95% confidence intervals.

The fact that the SRN is the more accurate psycholinguistic model does not mean that the PCFG has no additional value: The combination of SRN- and PCFG-based surprisal estimates may account for more of the variance in RT than do the SRN's estimates by themselves. To investigate whether this is the case, three linear mixed-effect regression models (see Baayen, Davidson, & Bates, 2008) were fitted. These three regression models differed only in the surprisal estimates that were included: Surprisal according to the PCFG (with $\theta = 16$), surprisal according to the SRN (after complete training), or both. In addition, the regression included fixed-effect factors for sentence length (number of characters), word frequency, forward transitional probability $\Pr(w_t|w_{t-1})$, and backward transitional probability $\Pr(w_t|w_{t+1})$. Also included were interactions between sentence length and backward transitional probability, and between word frequency and forward transitional probability.[4] All these factors had a significant effect, but no other two-way interactions did.

Table 1 shows the estimated β-coefficients for the effect of surprisal. When only PCFG-based or only SRN-based surprisal estimates are included, these have a significant positive effect on RT. More importantly, when both surprisal estimates are included, the effect of PCFG-based surprisal is no longer significant. This shows that these surprisal estimates do not add useful information to those generated by the SRN.

Another way to analyze the difference between the three regression models is by comparing model fits, as expressed by the Akaike Information Criterion (AIC; Akaike, 1974) in Table 1 (note that lower AIC indicates a better fitting model).

[4]Word frequencies and transitional probabilities were in fact log-transformed and averaged over all words of a sentence. The probabilities were estimated from word and bigram frequencies that were the average of relative frequencies in the Dundee corpus and in the written text part of the British National Corpus (BNC). This average of frequencies over two corpora turned out to correlate more strongly with word reading times than did frequencies from each corpus individually.

The only random-effect factors were by-subject and by-item intercepts, and a by-subject effect of word frequency. Including other random effects did not improve model fit significantly.

Table 1: Estimated coefficients (with associated p-values) of surprisal factors, and AIC of regression models.

Surprisal estimated by		Regression model includes PCFG	SRN	both
PCFG		$\beta = 0.45$		$\beta = -.46$
		$p < .02$		$p > .2$
SRN			$\beta = .64$	$\beta = 1.02$
			$p < .001$	$p < .01$
	AIC	104151	104145	104145

As was expected, such an analysis leads to the same conclusion: The regression model with both SRN- and PCFG-based surprisal estimates fits the RT-data better than the PCFG-only model ($\chi^2 = 7.54; p < .01$), whereas there is no significant difference in fit between the full model and the SRN-only model ($\chi^2 = 1.44; p > .2$).

Qualitative analysis To find out what causes the higher accuracy of the SRN's predictions, we first investigated whether it results from a slightly better fit to the RT data overall, or from a much better fit to a particular group of data points. Two ordinary linear regression models were fitted to the centered RTs (one including only the SRN-based surprisals, the other with only the PCFG-based surprisals) and their residuals were compared. If a particular data point is predicted more accurately by the SRN than by the PCFG, the corresponding residual in the SRN regression model will be closer to zero than that residual in the PCFG regression model. Hence, the extent to which data point i is predicted more accurately by the SRN than by the PCFG can be expressed by

$$\delta_i = |\text{resid}_i(\text{PCFG})| - |\text{resid}_i(\text{SRN})|,$$

where $|\text{resid}_i(\text{MODEL})|$ is the absolute residual of data point i in the regression using surprisals estimated by MODEL.

If the difference between SRN and PCFG were due solely to random noise, the δs would be distributed symmetrically around 0. In fact, the SRN fits the data better so the mean of δ is positive. If this is because of a particular group of data points, the corresponding δs would be exceptionally large, resulting in a distribution that is *skewed* to the right. Alternatively, if the SRN provides better fit in general, the distribution of δ would be *shifted* rightwards but remain symmetrical. Therefore, the question whether the SRN's better fit holds only for a subset of the data can be operationalized as: Are the δs distributed asymmetrically?

It turns out that they are not: A two-sample Kolmogorov-Smirnov test showed that the distribution is not significantly asymmetric ($p > .17$). Also, a scatter plot of $|\text{resid}_i(\text{PCFG})|$ against $|\text{resid}_i(\text{SRN})|$ did not show a single outlier, which would have indicated a data point predicted much better by one model than by the other. These findings suggests that the SRN does not just do better in particular cases, but is the more accurate psycholinguistic model overall.

Discussion

Several conclusions can be drawn from the results displayed in Figure 2. First, both PCFG- and SRN-based estimates of unlexicalized surprisal correlate positively with RT (provided beam width is large enough and the SRN is sufficiently trained). These correlations are very weak, but it should be kept in mind that many factors that affect reading times are not captured by surprisals of pos-tags (not to mention the presence of unexplainable noise in the data). For example, reading slows down on words that are infrequent or semantically unexpected in the context, but unlexicalized surprisal is not related to such factors (nor to many others). Moreover, the RT data were collected over general newspaper texts rather than experimental stimuli that are carefully constructed to balance as many factors as possible. For these reasons, we should not expect unlexicalized surprisal to explain much of the variance in RT. The fact that there is *any* significant, positive relation between unlexicalized surprisal and RT is enough to support the unlexicalized surprisal theory.

Second, it is likely that reading-time delays are actually caused by the unexpected occurrence of a part-of-speech, rather than there being some confounding variable responsible for this apparent effect. This is because such a confounding variable should also account for the relation between language model accuracy and psycholinguistic model accuracy. For example, the number of characters in a sentence has a (very weak) negative correlation with both RT[5] and surprisal. Sentence length might therefore be a confounding variable that is responsible for the relation between surprisal and RT. However, for the confound to also explain the slopes of the lines in Figure 2, it would need to correlate more strongly with surprisal as the language model becomes more accurate. There is no reason to believe that this would be the case for sentence length, and in fact it is not. Also, regression models that include (among others) a sentence-length factor showed a significant effect of surprisal on RT (see Table 1), which means that this effect does not depend on a confound with sentence length.

Third, the SRN seems to be a more accurate psycholinguistic model than the PCFG. This was confirmed by the regression analysis: the SRN explained RT-variance in addition to the PCFG, but the reverse was not the case. It was shown that this was not because the SRN does better than the PCFG on particular sentences. Rather, it is a more accurate psycholinguistic model overall.

Fourth, the monotonous relation between language model accuracy and psycholinguistic model accuracy indicates that readers make use of an objectively accurate language model. This was implicitly assumed in the research by Boston et al. (2008), Demberg and Keller (2008), Levy (2008), and Smith and Levy (2008), but has not been empirically validated before. It is noteworthy that this monotonous relation seems to hold *within* model type (SRN or PCFG) but not *between* these types: The PCFG is generally the more accurate lan-

[5] Be reminded that RT is defined as *per character* reading time.

guage model even though the SRN is the more accurate psycholinguistic model, suggesting that there is some important qualitative difference between them.

This raises the question what might make the SRN a better description of the human sentence-processing system. Although PCFGs (and tree-structure models in general) are particularly good at dealing with long-term dependencies within sentences, they do not directly store word sequences. SRNs, on the other hand, do retain information about frequencies of word sequences, but have difficulties with long-term dependencies. Possibly, people are more like SRNs than like PCFGs in this respect, at least insofar as is relevant for speed of reading. Indeed, experimental evidence indicates that children store frequent multi-word sequences as wholes (Bannard & Matthews, 2008) and that adults read frequent word sequences faster than less frequent ones (Bod, 2001; Tremblay, Derwing, Libben, & Westbury, 2008). The results presented here suggest that the same may be the case for part-of-speech sequences.

Conclusion

The finding that the SRN's predictions of reading-time data are more accurate than the PCFG's is of particular interest considering the ongoing debate about the nature of mental representations involved in human sentence processing. According to standard linguistic theories since Chomsky (1957), the mental representation of sentences involves syntactic tree structures. In contrast, connectionist theories of language processing take mental representations to be unstructured activation patterns. The fundamental difference between these two types of model has stood in the way of an objective and quantitative comparison of their ability to account for experimental data. Yet, such a comparison is precisely what was presented here. The result was that the connectionist model outperforms the symbolic model.

No matter how convincing this outcome may seem, it remains to be investigated to what extent it generalizes to other data sets, other instantiations of SRNs and PCFGs, and other models. The surprisal-based evaluation method may be particularly suited to such an investigation, because many (if not all) probabilistic sentence-processing models can provide surprisal estimates. Even models that differ widely from one another, such as symbolic and connectionist models, can thereby be evaluated against one and the same data set. Surprisal can act as a point of convergence for different models of sentence processing, allowing for a fair comparison between them.

Acknowledgments

I would like to thank Vera Demberg and Victor Kuperman for answering my many questions; Brian Roark for making his parser available (and answering even more questions); and four anonymous reviewers, Rens Bod, Gideon Borensztajn, and Victor again for their helpful comments. The research presented here was supported by grant 277-70-006 of the Netherlands Organization for Scientific Research (NWO).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

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, 59*, 390–412.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on childrens repetition of four-word combinations. *Psychological Science, 19*, 241–248.

Bod, R. (2001). *Sentence memory: Storage vs. computation of frequent sentences.* Paper presented at CUNY-2001, Philadelphia, PA.

Boston, M. F., Hale, J., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research, 2*, 1–12.

Brill, E. (1993). *A corpus-based approach to language learning.* Doctoral dissertation, University of Pennsylvania.

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton.

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

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

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

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

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

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

Santorini, B. (1991). *Part-of-speech tagging guidelines for the Penn Treebank project* (Tech. Rep.). Philadelphia, PA: University of Pennsylvania.

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 595–600). Austin, TX: Cognitive Science Society.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2008). *Processing advantages of lexical bundles: Evidence form self-paced reading experiments, word and sentence recall tasks, and off-line semantic ratings.* (Manuscript submitted for publication)