Research Article

Psychological Science
22(6) 829–834

DOI: 10.1177/0956797611409589
http://pss.sagepub.com

Insensitivity of the Human Sentence-Processing System to Hierarchical Structure

Stefan L. Frank and Rens Bod
Institute for Logic, Language and Computation, University of Amsterdam

Abstract

Although it is generally accepted that hierarchical phrase structures are instrumental in describing human language, their role in cognitive processing is still debated. We investigated the role of hierarchical structure in sentence processing by implementing a range of probabilistic language models, some of which depended on hierarchical structure, and others of which relied on sequential structure only. All models estimated the occurrence probabilities of syntactic categories in sentences for which reading-time data were available. Relating the models' probability estimates to the data showed that the hierarchical-structure models did not account for variance in reading times over and above the amount of variance accounted for by all of the sequential-structure models. This suggests that a sentence's hierarchical structure, unlike many other sources of information, does not noticeably affect the generation of expectations about upcoming words.

Keywords

cognitive processes, computer simulation, language, psycholinguistics, neural networks

Received 12/24/09; Revision accepted 2/3/11

Hierarchical phrase structures are considered fundamental to any description of the syntax of human languages because of their ability to handle nonadjacent, hierarchical dependencies between the words of a sentence (Chomsky, 1957; Hauser, Chomsky, & Fitch, 2002). Nevertheless, it is still unclear what role these structures play in the cognitive process of sentence comprehension. Although models based on phrase-structure grammars (PSGs) have appeared in explanations of several psycholinguistic findings (e.g., see Levy, 2008), it has also been argued that people's sensitivity to hierarchical structure is limited, as evidenced by, for instance, the ease with which some ungrammatical structures can be processed (Christiansen & MacDonald, 2009).

In the experiment reported here, we approached the issue differently: Rather than arguing that a particular linguistic (or psycholinguistic) phenomenon forms evidence for or against the use of hierarchical structures in sentence processing, we tested three types of probabilistic language models, each of which embedded different psychological mechanisms and representations. We then compared how well word-probability estimates generated by these different types of models accounted for a large set of reading-time measurements over general texts. If hierarchical structures, such as the one shown in Figure 1a, play a role in human sentence processing, then models that adopt them should fit the data better than models that do not. What we found, however, is that hierarchical-structure models did not generally fit the data better than sequential-structure models did. In fact, reading times were predicted more accurately by recurrent neural networks that use only sequential structure.

Method

In computational linguistics, a language model is defined as a probability model that estimates $P(w_t|w_{1\ldots t-1})$, which is the probability of the word occurring at sentence position t given the sentence's previous words $w_{1\ldots t-1}$ (Jurafsky & Martin, 2009). In our experiment, we used such probability estimates to account for word-reading times. We generated these estimates using language models of three different types: probabilistic PSGs, Markov models, and echo state networks (ESNs).[1] PSG models use hierarchical structures (see Fig. 1a), whereas Markov models and ESNs estimate probabilities that depend solely on the sentences' sequential structure (as shown in Fig. 1b).

Corresponding Author:
Stefan L. Frank, Department of Cognitive, Perceptual and Brain Sciences, University College London, 26 Bedford Way, London WC1H 0AP, United Kingdom
E-mail: s.frank@ucl.ac.uk

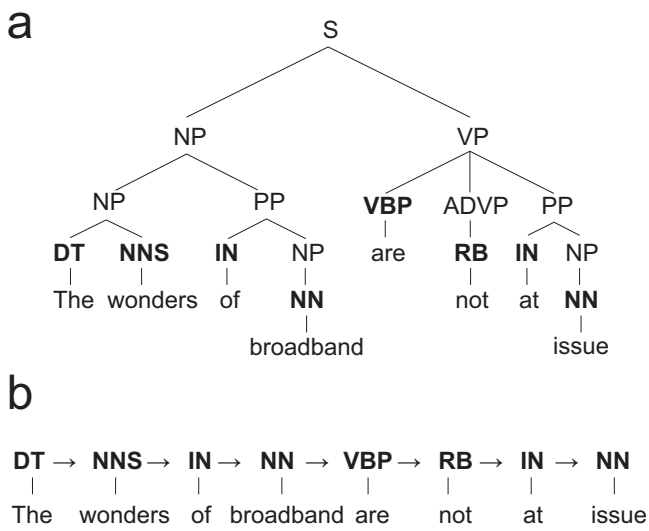

Fig. 1. Example (a) hierarchical phrase structure and (b) sequential structure for a sentence from the Dundee corpus (Kennedy & Pynte, 2005). Part-of-speech tags are shown in boldface (DT = determiner, NNS = plural noun, IN = preposition, NN = singular noun, VBP = present-tense verb, RB = adverb). Phrasal labels are shown in regular font (S = sentence, NP = noun phrase, PP = prepositional phrase, VP = verb phrase, ADVP = adverbial phrase).

Model-training data

All models used in our experiment were trained on the Wall Street Journal (WSJ) corpus of the Penn Treebank II (Marcus, Santorini, & Marcinkiewicz, 1993), which comprises 49,208 sentences annotated with syntactic tree structures. Such a collection of syntactically annotated sentences is known as a *treebank*.[2] Rather than using the corpus's actual words, we replaced each word with its part-of-speech (POS) tag. These tags correspond with the 45 syntactic categories in the treebank (e.g., "are" was classified as a present-tense verb, the tag for which is VBP; see Fig. 1). The probability of a syntactic category can be estimated much more accurately than the probability of a word can, and the former has been found to predict reading times more reliably than the latter (Demberg & Keller, 2008). An additional advantage of using POS tags instead of words is that tags have no lexical semantics, and this ensures that the results obtained using POS tags are indicative purely of linguistic structure. Because Markov models and ESNs use only sequential structure, they were trained on the sequences of POS tags only; that is, they ignored the sentences' tree structures.

Language models

PSG models. A PSG consists of a set of production rules and associated probabilities, which can be induced from a treebank. Each production rule corresponds with a parent node and its children in a tree structure. For example, in Fig. 1a, the PP (prepositional phrase) nodes are the parents of the children IN (preposition) and NP (noun phrase). In a standard probabilistic context-free grammar, the probability that a parent node will produce a particular set of children is conditional only on the parent node, but it is well known that parsing accuracy can be improved by taking the grandparent node (e.g., in Fig. 1a, VP, or verb phrase, is the grandparent of the rightmost IN and NP) into account (Johnson, 1998). Likewise, a rule's probability can be made conditional on information from even higher up in the parse tree (e.g., from a great-grandparent node, such as S, or sentence, which is the great-grandparent of all nodes that are three levels lower in the tree).

Using an algorithm that allows conditioning of the rule probabilities on any desired set of features from the parse trees (Roark, 2001), we obtained four different PSGs by varying the levels in the tree from which conditioning information was obtained: from only Level 1 (i.e., a standard probabilistic context-free grammar) up to Level 4, at most. In addition, we induced four more PSGs, in which conditioning information was taken not only from ancestor nodes (e.g., grandparent nodes) but also from the ancestors' left siblings (e.g., in Fig. 1a, the left sibling of VP is NP), again varying the maximum number of levels up in the tree from one to four. In this manner, we obtained highly structurally sensitive syntactic models. We refer to the two types of grammars as *PSG-a* (using only ancestor information) and *PSG-s* (taking also the ancestors' left siblings into account). In addition, probabilities were also conditioned on the current head node, and this made the grammars sensitive to subject-verb number agreement.[3]

An incremental (i.e., word-by-word) parser (Roark, 2001) was used to obtain the desired probability estimates for $P(w_t|w_{1\ \ldots\ t-1})$ generated by each grammar. Fig. 1a shows the most probable parse of the POS tags of one example sentence according to five of our eight PSG models.

Markov models. We used Markov models of first, second, and third orders (ns = 1, 2, and 3, respectively) in our experiments. In a Markov model of a given order (n), a symbol's probability depends on the context of n previous symbols only; that is, $P(w_t|w_{1\ \ldots\ t-1})$ is taken to equal $P(w_t|w_{t-n\ \ldots\ t-1})$. In our experiments, the probabilities of the sequences $w_{t-n\ \ldots\ t-1}$ and $w_{t-n\ \ldots\ t}$ were estimated from their occurrence frequencies in the training data. The raw frequency counts were smoothed because many longer sequences were rare or even absent from the data. Because the particular smoothing method affects model accuracy, we used three different methods: additive smoothing, Simple Good-Turing smoothing (Gale & Sampson, 1995), and Witten-Bell smoothing (Witten & Bell, 1991). For first-order models, only additive smoothing was applied because differently smoothed first-order models are nearly identical to one another. For second- and third-order models, all three smoothing methods were used. Thus, seven Markov models were created in total.

ESNs. ESNs (Jaeger & Haas, 2004) are recurrent neural networks that can be trained more efficiently than the more

common simple recurrent networks (SRNs; Elman, 1991) because the weights of an ESN's input and recurrent connections remain fixed at random values. Consequently, output-weight training reduces to a linear regression of the target outputs on the transient activations of the ESN's recurrent-layer units. In our experiments, we used ESNs rather than SRNs because it has been argued that SRNs can learn the hierarchical structure of language through adjustment of the recurrent connection weights (Elman, 1991). This is clearly not the case with ESNs, as their recurrent connection weights are never adjusted. Therefore, ESNs can use only the sentences' sequential structure, as do Markov models. In contrast with Markov models, however, there is no upper limit to the length of the available context, making ESNs more plausible as cognitive models of sentence processing.

ESNs have successfully been applied to next-symbol prediction in sentence processing (Frank & Čerňanský, 2008; Tong, Bickett, Christiansen, & Cottrell, 2007). In our experiments, we trained six ESNs on the POS-tag sequences of the WSJ treebank. The number of hidden units in these networks varied from 100 to 600. The networks' output-activation vectors were interpreted as probability distributions over possible upcoming POS tags. That is, the activation of the output unit corresponding to the actual next POS tag w_t was used as the desired estimate of the probability $P(w_t|w_{1\ldots t-1})$.

Model evaluation

Evaluation data set. The models were tested on the same data set used in several earlier studies (Demberg & Keller, 2008; Frank, 2009; Smith & Levy, 2008): the Dundee corpus (Kennedy & Pynte, 2005), which consists of 2,368 sentences from British newspaper editorials. The Dundee corpus comprises 51,501 word tokens, which were assigned POS tags[4] in accordance with the Penn Treebank guidelines (Santorini, 1991). The corpus comes with eye-tracking data from 10 participants, from which we extracted three different measures of word-reading time: the first-pass, right-bounded, and go-past durations.

Following Demberg and Keller (2008), we removed data points (i.e., word-participant pairs) if the word was not fixated, was presented as the first or last on a line, was attached to punctuation, contained more than one capital letter, or contained a nonletter (this included clitics, which Demberg and Keller did not remove). Mainly because of the large number of nonfixations (more than 46%), 62.8% of data points were removed, leaving 191,380 data points (between 16,469 and 21,770 per participant).

Linking model outcomes with reading times. Fundamental differences between types of sentence-processing models have often hampered a direct comparison of their abilities to account for empirical findings. For example, Keller (2003) suggests that sentence-acceptability judgments can be predicted by the probability of the sentence's most probable parse. However, such a measure is not available for sentence-processing models that do not construct parses, such as Markov models and most connectionist models. Alternatively, the measure for word-prediction error proposed by Christiansen and Chater (1999) to account for processing difficulty can be computed only if the true probabilities $P(w_t|w_{1\ldots t-1})$ are known. This is the case in simulations using artificial, predefined, miniature languages (as is typical for much connectionist research), but when dealing with natural text corpora, as we did in this experiment, there is no such thing as the true probability of a word or POS tag.

We solved this problem by using surprisal theory (Hale, 2001; Levy, 2008) to link model outcomes with reading times. Informally, surprisal theory simply states that more cognitive effort is required to process words whose occurrence is less expected. More specifically, if $w_{1\ldots t-1}$ denotes the $t-1$ words that have occurred in a sentence so far, the amount of cognitive effort needed to process the next word, w_t, is positively linearly related to $-\log P(w_t|w_{1\ldots t-1})$, a value known as the *surprisal* of w_t. Because the probabilities $P(w_t|w_{1\ldots t-1})$ may be estimated by any probabilistic language model, surprisal estimates can be used to compare different types of models.

Surprisal theory can be derived from different assumptions about human sentence processing (Levy, 2008; Smith & Levy, 2008) and indeed the relation between word surprisal (as estimated by different types of models) and processing effort (as reflected in reading times) has been confirmed in several studies (Boston, Hale, Patil, Kliegl, & Vasishth, 2008; Demberg & Keller, 2008; Frank, 2009; Roark, Bachrach, Cardenas, & Pallier, 2009; Smith & Levy, 2008; Wu, Bachrach, Cardenas, & Schuler, 2010).

Evaluation measures. Each model was evaluated on two measures: linguistic accuracy and psychological accuracy. The first, which indicates how well the model has captured patterns in the language, is defined as the negative of the model's average surprisal estimated over the test corpus. The higher (i.e., less negative) this value, the "less surprised" the model is by the test corpus, meaning that it forms a more accurate model of the language.

Psychological accuracy indicates how well the model captures patterns in the reading-time data. This was determined by first fitting to the data a mixed-effects regression model (Baayen, Davidson, & Bates, 2008) with several predictors that are known to account for word-reading time, such as word length and word frequency.[5] This regression model did not contain any of the surprisal estimates. Next, the set of surprisal estimates generated by one of the language models was included in the regression. The resulting decrease in the regression model's deviance was indicative of the amount of variance in reading time accounted for by those surprisal estimates, and this decrease was taken as the measure for psychological accuracy of the model that generated those estimates.[6]

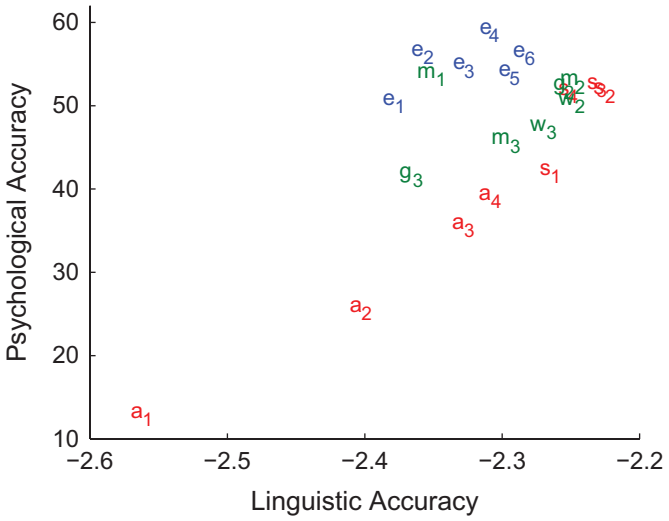

Fig. 2. Results for first-pass reading times: psychological accuracy plotted against linguistic accuracy. Psychological accuracy was defined as the decrease in deviance that resulted from including one set of surprisal estimates in the regression, and linguistic accuracy was defined as the negative of the average surprisal. Results are plotted for phrase-structure-grammar (PSG) models, Markov models, and echo state networks. PSG models were constructed using only ancestor information (a_n, where n indicates the number of levels up in the parse tree from which conditioning information was obtained) or taking also the ancestors' left siblings into account (s_n). Markov models of order n were created with additive smoothing (m_n), Simple Good-Turing smoothing (g_n), or Witten-Bell smoothing (w_n). Echo state networks (e_n) had 100n hidden units.

Results

Results for first-pass durations[7] are presented in this article; the other two sets of results can be found in the Supplemental Material available online. Figure 2 shows each model's linguistic accuracy plotted against its psychological accuracy. Each set of surprisal estimates contributes significantly, all χ^2s(1, N = 191,380) > 12.8; $p < .0004$, and in the correct (i.e., positive) direction to the regression model's fit to first-pass reading times.

The PSG models were able to reach higher levels of linguistic accuracy than Markov models and ESNs were.[8] Moreover, there was a clear relation between the PSG models' linguistic and psychological accuracies: More accurate models of the language also predicted the reading times more accurately. The same relation seems to hold, albeit not as strongly, for the sequential-structure models. A comparison between model types, however, showed that, at similar levels of linguistic accuracy, the ESNs had higher psychological accuracy than did the PSG models. The psychological accuracy of Markov models is either above or equal to that of PSG models with similar linguistic accuracy.

ESNs formed more accurate psychological models than PSGs did; however, this does not mean that hierarchical structure lacks the ability to account for any unique variance in reading time. To investigate whether hierarchical structure had additional explanatory value, we compared the ESN and PSG that showed highest psychological accuracy (i.e., the 400-unit ESN and Level 3 PSG-s) by taking the regression model that includes either the PSG model's or the ESN's surprisal estimates and adding the surprisal estimates generated by the other language model. The resulting decreases in deviance revealed that the PSG model's estimates did not significantly contribute to the estimates made by the ESN, $\chi^2(1, N = 191{,}380) = 0.95$; $p > .3$, whereas the ESN-based surprisals do have predictive value over and above the PSG model's, $\chi^2(1, N = 191{,}380) = 7.56$; $p < .006$. This shows that the PSG does not explain variance in reading-time data over and above what is already accounted for by the ESN. Consistent results were obtained using the two alternative reading-time measures (see the Supplemental Material for details).

Discussion

The best-performing PSG models were more linguistically accurate than Markov models and ESNs were. Nevertheless, having access to hierarchical phrase structure did not always make PSG models psychologically more accurate than models that use only sequential structure. On the contrary, ESNs, which do not adopt hierarchical structure, estimated surprisal values that fit the reading times better than PSG models did. This finding suggests that human sentence processing relies more on sequential than on hierarchical structure, at least insofar as is relevant for generating expectations about upcoming material. It should be kept in mind, however, that language models (and in particular hierarchical ones) come in many more varieties than the selection we have studied here. It remains to be investigated whether the current results generalize to a wider set of sequential and hierarchical language models.

Nonadjacent dependencies are ubiquitous in language and many appear in the Dundee corpus. The sentence displayed in Figure 1 is an example: The plural verb "are" is dependent on the plural noun "wonders" and not on the adjacent singular noun "broadband." PSG models are particularly good at dealing with such nonadjacent, long-term dependencies within sentences (Chomsky, 1957; Manning & Schütze, 1999) but do not directly store word or POS sequences. In contrast, Markov models and ESNs do retain information about frequencies of sequences, but have difficulties with long-term dependencies. Possibly, people behave more like ESNs than like PSGs in this respect. Indeed, experimental evidence has provided at least five indications of this possibility: Frequent multiword sequences are stored as wholes by both children (Bannard & Matthews, 2008) and adults (Arnon & Snider, 2010), more frequent word sequences are read faster than less frequent ones (Tremblay, Derwing, Libben, & Westbury, 2011), locally coherent structure can interfere with long-term dependencies (Tabor, Galantucci, & Richardson, 2004), sensitivity to sequential structure is correlated with sensitivity to word predictability (Conway, Bauernschmidt, Huang, & Pisoni, 2010), and subject-verb number-agreement errors in sentence production depend on the sentence's sequential rather than hierarchical structure (Gillespie & Pearlmutter, 2011).

It does not directly follow from our findings that hierarchical structure plays no role whatsoever in sentence processing. In fact, experimental evidence suggests that such structures may be relevant in specific cases. For example, Staub and Clifton (2006) found that the occurrence of "either" speeds up reading of the words following the corresponding "or." Using a hierarchical-structure model, Demberg and Keller (2009) provided a surprisal-based explanation of this finding, but it remains to be shown that a model that uses only sequential structure will not suffice.

Although our results are not informative regarding the processing of specific constructions such as "either . . . or," they do indicate that, in general, a sentence's hierarchical structure does not measurably affect readers' expectations about the next input. This is especially noticeable because so many other sources of information have been shown to be efficacious, from simple bigram (McDonald & Shillcock, 2003) and multiword statistics (Tremblay et al., 2011) to information from prior discourse (Otten & Van Berkum, 2008), nonlinguistic context (Kamide, Altmann, & Haywood, 2003), and world knowledge (Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005). Therefore, if phrase-structure information were available during reading, we would also expect it to affect readers' expectations about upcoming words. Instead, these expectations seem unaffected by the sentence's hierarchical structure.

Acknowledgments

We would like to thank Philémon Brakel, Victor Kuperman, Roger Levy, Brian Roark, and two anonymous reviewers for their help and comments.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

The research presented here was funded by the Netherlands Organization for Scientific Research (NWO; Grant Number 277-70-006) and by the European Union 7th Framework Programme (FP7/2007-2013; Grant Number 253803).

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Notes

1. The technical details of these models can be found in the Supplemental Materials available online.
2. Syntactic trees were preprocessed in a standard manner (Manning & Schütze, 1999), which involved removing traces and semantic information on phrase labels.
3. The head of an NP is its noun, so, using the nodes in the tree depicted in Figure 1a as an example, before the parser expands the VP, the current head node is a plural noun (NNS). If a production rule's probability is conditioned on the current head node, the VP is therefore more likely to expand to a plural verb than to a singular verb.
4. The POS tags were taken from Frank (2009). Tagging was done automatically using Brill's (1993) POS tagger, after which all tags were checked by hand.
5. See the Supplemental Material for details of the regression analysis.
6. The decrease in deviance equals –2 times the log-likelihood ratio of the two regression models and follows an approximate chi-square distribution with 1 degree of freedom.
7. A first-pass duration is the total duration of all fixations on a word until the first fixation on any other word.
8. A Wilcoxon matched-pairs signed-rank test comparing the linguistically most accurate models of each type showed that the Level 2 PSG has significantly higher linguistic accuracy (compared with the second-order Markov model: $z = 15.6$, $p \approx 0$; compared with the 600-unit ESN: $z = 21.7$; $p \approx 0$). For these tests, linguistic accuracies were averaged over the POS tags of each sentence to avoid artifacts resulting from dependencies between the surprisal values within a sentence.

References

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82.

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 children's repetition of four-word combinations. *Psychological Science*, *19*, 241–248.

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). Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1193&context=ircs_reports&sei-redir=1#

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

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205.

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, *59*, 129–164.

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*, 356–371.

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

Demberg, V., & Keller, F. (2009). A computational model of prediction in human parsing: Unifying locality and surprisal effects.

In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1888–1893). Austin, TX: Cognitive Science Society.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1139–1144). Austin, TX: Cognitive Science Society.

Frank, S. L., & Čerňanský, M. (2008). Generalization and systematicity in echo state networks. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 733–738). Austin, TX: Cognitive Science Society.

Gale, W. A., & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, *2*, 217–237.

Gillespie, M., & Pearlmutter, N. J. (2011). Hierarchy and scope of planning in subject-verb agreement production. *Cognition*, *118*, 377–397.

Hale, J. (2001). A probabilistic Earley 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.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, *298*, 1569–1579.

Jaeger, H., & Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, *304*, 78–80.

Johnson, M. (1998). The effect of alternative tree representations on tree bank grammars. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning* (pp. 39–48). Pittsburgh, PA: Association for Computational Linguistics.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing*. Upper Saddle River, NJ: Pearson Education.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*, 133–156.

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: Cognitive Science Society.

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.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

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

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.

Otten, M., & Van Berkum, J. J. A. (2008). Discourse-based word anticipation during language processing: Prediction or priming? *Discourse Processes*, *45*, 464–496.

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

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 324–333). Pittsburgh, PA: Association for Computational Linguistics.

Santorini, B. (1991). *Part-of-speech tagging guidelines for the Penn Treebank Project* (Technical Report No. MS-CIS-90-47). Philadelphia: 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.

Staub, A., & Clifton, C., Jr. (2006). Syntactic prediction in language comprehension: Evidence from *either . . . or*. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 425–436.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*, 355–370.

Tong, M. H., Bickett, A. D., Christiansen, E. M., & Cottrell, G. W. (2007). Learning grammatical structure with echo state networks. *Neural Networks*, *20*, 424–432.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*. Advance online publication. doi:10.1111/j.1467-9922.2010.00622.x

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 443–467.

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

Wu, S., Bachrach, A., Cardenas, C., & Schuler, W. (2010). Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 1189–1198). Pittsburgh, PA: Association for Computational Linguistics.