Evaluating Models of Computation and Storage in Human Sentence Processing

Minh-Thang Luong*
Stanford University
lmthang@stanford.edu

Timothy J. O'Donnell*
MIT
timod@mit.edu

Noah D. Goodman
Stanford University
ngoodman@stanford.edu

Abstract

We examine the ability of several models of computation and storage to explain reading time data. Specifically, we demonstrate on both the Dundee and the MIT reading time corpora, that fragment grammars, a model that optimizes the tradeoff between computation and storage, is able to better explain people's reaction times than two baseline models which exclusively favor either storage or computation. Additionally, we make a contribution by extending an existing incremental parser to handle more general grammars and scale well to larger rule and data sets.[1]

1 Introduction

A basic question for theories of language representation, processing, and acquisition is how the linguistic system balances storage and reuse of lexical units with productive computation. At first glance, the question appears simple: words are stored; phrases and sentences are computed. However, a closer look quickly invalidates this picture. Some canonically computed structures, such as phrases, must be stored, as witnesses by verbal idioms like *leave no stone unturned*[2] (Nunberg et al., 1994). There is also compositionality at the sub-word level: affixes like *ness* in *pine-scentedness*, are almost always composed productively, whereas other affixes, e.g., *th* in *warmth*, are nearly always stored together with stems (O'Donnell, 2015). Facts such as these have led to a consensus in the field that storage and computation are properties that cut across different kinds of linguistic units and levels of linguistic structure (Di Sciullo and Williams, 1987)—giving rise to *hetergeneous lexicon*[3] theories, in the terminology of Jackendoff (2002b).

Naturally, the question of what is computed and what is stored has been the focus of intense empirical and theoretical research across the language sciences. On the empirical side, it has been the subject of many detailed linguistic analyses (e.g., Jackendoff (2002a)) and specific phenomena such as composition versus retrieval in word or idiom processing have been examined in many studies in experimental psycholinguistics (Hay, 2003; O'Donnell, 2015). On the theoretical side, there have been many proposals in linguistics regarding the structure and content of the heterogeneous lexicon (e.g., Fillmore et al. (1988), Jackendoff (2002b)). More recently, there have been a number of proposal from computational linguistics and natural language processing for how a learner might infer the correct pattern of computation and storage in their language (De Marcken, 1996; Bod et al., 2003; Cohn et al., 2010; Post and Gildea, 2013; O'Donnell, 2015).

However, there remains a gap between detailed, phenomenon-specific studies and broad architectural proposals and learning models. Recently, however, a number of methodologies have emerged which promise to bridge this gap. These methods make use of broad coverage probabilistic models which can encode representational and inferential assumptions, but which can also be applied to make detailed predictions on large psycholinguistic datasets encompassing a wide vari-

*indicates equal contribution.

[1] Our code and data are available at http://stanford.edu/~lmthang/earleyx/.

[2] Meaning: *prevent any rock from remaining rightside up*.

[3] A hetergeneous lexicon contains not only words but also affixes, stems, and phrasal units such as idioms.

ety of linguistic phenomena. In the realm of syntax, one recent approach has been to use probabilistic models of sentence structures, paired with incremental parsing algorithms, to produce precise quantitative predictions for variables such as reading times (Roark et al., 2009) or eye fixation times (Demberg and Keller, 2008; Mitchell et al., 2010; Frank and Bod, 2011; Fossum and Levy, 2012; van Schijndel and Schuler, 2013). To date, no models of storage and computation in syntax have been applied to predict measures of human reading difficulty.

In this work, we employ several of the models of computation and storage studied by O'Donnell (2015), to examine human sentence processing. We demonstrate that the fragment grammars model (O'Donnell et al., 2009; O'Donnell et al., 2011)—a model that treats the question of what to store and what to compute productively as a probabilistic inference—better explains human reading difficulty than two "limiting-case" baselines, MAP adaptor grammars (maximal storage) and Dirichlet-multinomial PCFG (maximal computation), in two datasets: the Dundee eye-tracking corpus (Kennedy and Pynte, 2005) and the MIT reading time dataset (Bachrach et al., 2009).

2 Goals and Scope of the Paper

Before moving on, we remark on the goals and scope of the current study. The emergence methods connecting wide-coverage probabilistic grammars and psycholinguistic data offer great potential to test theoretical models quantitatively, at scale, and on a variety of detailed phenomena. However, studies using these methods also involve many moving parts, often making their results difficult to interpret.

To connect probabilistic models of syntactic computation and storage to reading time or eye fixation data, practioners need to:

1. Preprocess train and test data sets by tokenizing words, limiting sentence lengths, and handling unknown words.

2. Decide on a suitable grammatical formalism: determine a hypothesis space of stored items and specify a probability model over that space.

3. Choose and implement a probabilistic model to extract grammars from the training set.

4. Pick a test set annotated with reading difficulty information, e.g., eye fixation or reading times.

5. Choose a specific incremental parsing algorithm to generate word-by-word parsing predictions.

6. Determine the theoretical quantity that will be used as a predictor, e.g., *surprisal* or *entropy reduction*.

7. Choose a suitable linking model to regress theoretical predictions against human data, controlling for participant-specific factors and nuisance variables.

Given this wide array of design decisions, it is often difficult to compare results across studies or to determine which theoretical assumptions are crucial to the performance of models. For the field to make progress, studies must be replicable and each of the above factors (and potentially others) must be varied systematically in order to isolate their specific consequences. We contribute towards this process in three ways.

First, we report results for three models which differ only in terms of how they address the problem of what to store and what to compute (see Section 3). Otherwise, modeling and analysis assumptions are exactly matched. Moreover, the models represent three "limiting cases" in the space of storage and computation — store all maximal structures, store only minimal structures, and treat the problem as a probabilistic inference. Although none of the models represents a state-of-the-art model of syntactic structure, this study should provide important baselines against which to compare in future proposals.

Second, to make this study possible, we extend an existing incremental parser to address two technical challenges by: (a) handling more general input grammars and (b) scaling better to extremely large rule sets. This parser can be used with any model that can be projected to or approximated by a probabilistic context-free grammar. We make this parser available to the community for future research.

Third, and finally, unlike previous studies which only report results on a single dataset, we demonstrate consistent findings over two popular datasets, the Dundee eye-tracking corpus and the MIT reading times corpus. We make available our

predicted values for all examined data points together with our analysis scripts. This should facilitate the replication of these specific results and direct numerical comparison with later proposals.

3 Approaches to Computation and Storage

In this paper we study the ability of three models to predict reading difficulty as measured by either eye-fixation or reading times — the *full-parsing* model, implemented by Dirichlet-multinomial probabilistic context-free grammars (DMPCFG) (Kurihara and Sato, 2006; Johnson et al., 2007), the *full-listing* mode, implemented by maximum a posteriori adaptor grammars (MAG) (Johnson et al., 2006), and the *inference-based* model, implemented by fragment grammars (FG) (O'Donnell, 2015).

All three models start with the same underlying *base system*—a context-free grammar (CFG) specifying the space of possible syntactic derivations—and the same training data—a corpus of syntactic trees. However, the models differ in what they store and what they compute. The full-parsing model can be understood as a fully-compositional baseline equivalent to a Bayesian version of the underlying CFG. The full-listing model, by contrast, stores all full derivations (i.e., all derivations down to terminal symbols) and subderivations in the input corpus. These stored (sub)trees can be thought of as extending the CFG base component with rules that directly rewrite nonterminal symbols to sequence of terminals in a single derivational step.

Finally, the inference-based model treats the problem of what tree fragments to store, and which parts of derivations to compute as an inference in a Bayesian framework, learning to store and and reuse those subtrees which best explain the data while taking into account two prior biases for simplicity. The first bias prefers to explain the data in terms of a smaller lexicon of stored tree fragments. The second bias prefers to account for each input sentence with smaller numbers of derivational steps (i.e., fragments). Note that these two biases compete and thus give rise to a tradeoff. Storing smaller, more abstract fragments allows the model to represent the input with a more compact lexicon, at the cost of using a greater number of rules, on average, in individual derivations. Storing larger, more concrete fragments allows the model to derive individual sentences using a smaller number of steps, at the cost of expanding the size of the stored lexicon. The inference-based model can be thought of as extending the base CFG with rules, inferred from the data, that expand larger portions of derivation-tree structure in single steps, but can also include nonterminals on their right-hand side (unlike the full-listing model).

As we mentioned above, none of these models take into account various kinds of structure—such as headedness or other category-refinements—that are known to be necessary to achieve state-of-the-art syntactic parsing results (Petrov et al., 2006; Petrov and Klein, 2007). However, the results reported below should be useful for situating and interpreting the performance of future models which do integrate such structure. In particular, these results will enable ablation studies which carefully vary different representational devices.

4 Human Reading Time Prediction

To understand the effect of different approaches to computation and storage in explaining human reaction times, we employ the surprisal theory proposed by Hale (2001) and Levy (2008). These studies introduced *surprisal* as a predictor of the difficulty in incremental comprehension of words in a sentence. Because all of the models described in the last section can be used to compute surprisal values, they can be used to provide predictions for processing complexity and hence, gain insights about the use of stored units in the human sentence processing. The surprisal values for these different models are dervied by means of a probabilistic, incremental Earley parser (Stolcke, 1995; Earley, 1968), which we describe below.

4.1 Surprisal Theory

The surprisal theory of incremental language processing characterizes the lexical predictability of a word w_t in terms of a surprisal value, the negative log of the conditional probability of a word given its preceding context, $-\log P(w_t|w_1 \dots w_{t-1})$. Higher surprisal values mean smaller conditional probabilities, that is, words that are less predictable are more surprising to the language user and thus harder to process. Surprisal theory was first introduced in Hale (2001) and studied more extensively by Levy (2008). It has also been shown to have a strong correlation with reading

time duration in both eye-tracking and self-paced reading studies (Boston et al., 2008; Demberg and Keller, 2008; Roark et al., 2009; Frank, 2009; Wu et al., 2010; Mitchell et al., 2010).

4.2 The Incremental Parser

The computation of surprisal values requires access to an incremental parser which can compute the *prefix probabilities* associated with a string s under some grammar—the total probability over all derivation using the grammar which generate strings prefixed by s (Stolcke, 1995). The prefix probability is an important concept in computational linguistics because it enables probabilistic predictions of possible next words (Jelinek and Lafferty, 1991) via the conditional probabilities $P(w_t|w_1 \dots w_{t-1}) = \frac{P(w_1 \dots w_t)}{P(w_1 \dots w_{t-1})}$. It also allows estimation of incremental costs in a stack decoder (Bahl et al., 1983). Luong et al. (2013) used prefix probabilities as scaling factors to avoid numerical underflow problems when parsing very long strings.

We extend the implementation by Levy (2008) of the probabilistic Earley parser described in Stolcke (1995) which computes exact prefix probabilities. Our extension allows the parser (a) to handle arbitrary CFG rewrite rules and (b) to scale well to large grammars.[4]

The implementation of Levy (2008) only extracts grammars implicit in treebank inputs and restricts all pre-terminal rules to single-terminal rewrites. To approximate the incremental predictions of the models in this paper, we require the ability to process rules that include sequences of multiple terminal and non-terminal symbols on their right-hand side. Thus, we extend the implementation to allow efficient processing of such structures (property a).

With regards to property (b), we note that parsing against the full-listing model (MAG) is prohibitively slow because the approximating grammars for the model contain PCFG rules which exhaustively list the mappings from every nonterminal in the input corpus to its terminal substring, leading to thousands of rules. For example, for the Brown corpus section of the Penn Treebank (Marcus et al., 1993), we extracted 778K rules for the MAG model, while the number of rules in the DMPCFG and the inference-based (FG) grammars are 75K and 146K respectively. Parsing the MAG is also memory intensive due to multi-terminal rules that rewrite to long sequences of terminals, because, for example, an S node must rewrite to an entire sentence. Such rules result in an exploding number of states during parsing as the Earley dot symbol moves from left to right.

To tackle this issue, we utilize a *trie* data structure to efficiently store multi-terminal rules and quickly identify (a) which rules rewrite to a particular string and (b) which rules have a particular prefix.[5] These extensions allow our implementation to incorporate multi-terminal rules in the prediction step of the Earley algorithm, and to efficiently incorporate which of the many rules can contribute to the prefix probability in the Earley scanning step.

We believe that our implementation should be useful to future studies of reading difficulty, allowing efficient computation of prefix probabilities for any model which can be projected to (or approximated by) a PCFG—even if that approximation is very large.

5 Experiments

5.1 Data

Our three models models are trained on the Wall Street Journal (WSJ) portion of the Penn Treebank (Marcus et al., 1994). In particular, because we have access to gold standard trees from this corpus, it is possible to compute the exact maximum a posteriori full-parsing (DMPCFG) and full-listing (MAG) models, and output PCFGs corresponding to these models.[6]

We evaluate our models on two different corpora: (a) the *Dundee corpus* (Kennedy and Pynte, 2005) with eye-tracking data on naturally occurring English news text and (b) the *MIT corpus* (Bachrach et al., 2009) with self-paced reading data on hand-constructed narrative text. The for-

[4] Other recent studies of human reading data have made use of the parser of Roark (2001). However, this parser incoporates many specific design decisions and optimizations—"baking in" aspects of both the incremental parsing algorithm and a model of syntactic structure. As such, since it does not accept arbitrary PCFGs, it is unsuitable for this present study.

[5] Specifically, terminal symbols are used as keys in our trie and at each trie node, e.g., corresponding to the key sequence a b c, we store two lists of nonterminals: (a) the *complete* list – where each non-terminal X corresponds to a multi-terminal rule $X \rightarrow a\ b\ c$, and (b) the *prefix* list – where each non-terminal X corresponds to a multi-terminal rule $X \rightarrow a\ b\ c \dots d$. We also accumulated probabilities for each non-terminal in these two lists as we traverse the trie.

[6] Note that for DMPCFG, this PCFG is exact, whereas for MAG, it represents a truncated approximation.

mer has been a popular choice in many sentence processing studies (Demberg and Keller, 2008; Mitchell et al., 2010; Frank and Bod, 2011; Fossum and Levy, 2012; van Schijndel and Schuler, 2013). The latter corpus, with syntactically complex sentences constructed to appear relatively natural, is smaller in size and has been used in work such as Roark et al. (2009) and Wu et al. (2010). We include both corpora to demonstrate the reliability of our results.

Detailed statistics of these corpora are given in Table 1. The last column indicates the number of data points (i.e., word-specific fixation or reading times) used in our analyses below. This dataset was constructed by excluding data points with zero reading times and removing rare words (with frequencies less than 5 in the WSJ training data). We also exclude long sentences (of greater than 40 words) for parsing efficiency reasons.

	sent	**word**	**subj**	**orig**	**filtered**
Dundee	2,370	58,613	10	586,131	228,807
MIT	199	3,540	23	81,420	69,702

Table 1: **Summary statistics of reading time corpora** – shown are the number of sentences, words, subjects, data points before (*orig*) and after filtering (*filtered*).

5.2 Metrics

Following (Frank and Bod, 2011; Fossum and Levy, 2012), we present two analyses of the surprisal predictions of our models: (a) a *likelihood* evaluation and (b) a *psychological* measure of the ability of each model to predict reading difficulty.

For the former, we simply average the negative surprisal values, i.e., $\log p(w_n|w_1 \ldots w_{n-1})$, of all words in the test set, computing the average log likelihood of the data under each model.[7] This can be understood as simply a measure of goodness of fit of each model on each test data set.

For the latter, we perform a linear mixed-effects analysis (Baayen et al., 2008) to evaluate how well the model explains reading times in the test data. The `lme4` package (Bates et al., 2011) is used to fit our linear mixed-effects models. Following (Fossum and Levy, 2012), eye fixation and reading times are log-transformed to produce more normally distributed data.[8] We include the following common predictors as fixed effects for each word/participant pair: (i) position of the word in the sentence, (ii) the number of characters in the word, (iii) whether the previous word was fixated, (iv) whether the next word was fixated, and (v) the log of the word unigram probability.[9]

All fixed effects were centered to reduce collinearity. We include by-word and by-subject intercepts as random effects. The *base* model results reported below include only these fixed and random factors. To test the ability of our three theoretical models of computation and storage to explain the reading time data, we include surprisal predictions from each model as an additional fixed effect. To test the signficance of these results, we perform nested model comparisons with χ^2 tests.

5.3 Results

For the *likelihood* evaluation, the values in Table 2 demonstrate that the FG model provides the best fit to the data. The results also indicate a ranking over the three models, FG $\succ$ DMPCFG $\succ$ MAG.

	Dundee	**MIT**
DMPCFG	-6.82	-6.80
MAG	-6.91	-6.95
FG	-6.35	-6.35

Table 2: **Likelihood Evaluation** – the average negative suprirsal values given by each model (DMPCFG, MAG, FG) on all words in each corpus (Dundee, MIT).

For the *psychological* evaluation, we present results of our nested model comparisons under two settings: (a) *additive* in which we independently add each of the surprisal measures to the *base* model and (b) *subtractive*, in which we take the *full* model consisting of all the surprisal measures and independently remove one surprisal measure each time.

Results of the additive setting are shown in Table 3, demonstrating the same trend as observed in the likelihood evaluation. In particular, the FG model yields the best improvement in terms of model fit as captured by the $\chi^2(1)$ statistics, indicating that it is more explanatory of reaction times when added to the *base* model as compared to the DMPCFG and the MAG predictions. The ranking

[7] Exponentiating this value gives the perplexity score.

[8] For the Dundee corpus, we use the first-pass reading time.

[9] The unigram probability was estimated from the WSJ training data, the written text portion of the BNC corpus, and the Brown corpus. We make use of the SRILM toolkit (Stolcke, 2002) for such estimation.

is also consistent with the likelihood results: FG $\succ$ DMPCFG $\succ$ MAG.

Models	Dundee		MIT	
	$\chi^2(1)$	p	$\chi^2(1)$	p
base+DMPCFG	70.9	$<$ 2.2E-16	38.5	5.59E-10
base+MAG	10.9	9.63E-04	0.1	7.52E-01
base+FG	118.3	$<$ 2.2E-16	62.5	2.63E-15

Table 3: **Psychological accuracy, additive tests** – $\chi^2(1)$ and p values achieved by performing nested model analysis between the models *base+X* and the *base* model.

For the subtractive setting, results in Table 4 highlight the fact that several models significantly ($p < 0.01$) explains variance in fixation times above and beyond the other surprisal-based predictors. The FG measure proves to be the most influential predictor (with $\chi^2(1) = 62.5$ for the Dundee corpus and 42.9 for the MIT corpus). Additionally, we observe that DMPCFG does not significantly explain more variance over the other predictors. This, we believe, is partly due to the presence of the FG model, which captures much of the same structure as the DMPCFG model.

Models	Dundee		MIT	
	$\chi^2(1)$	p	$\chi^2(1)$	p
full-DMPCFG	4.0	4.65E-02	3.5	6.18E-02
full-MAG	14.3	1.58E-04	23.6	1.21E-06
full-FG	62.5	2.66E-15	42.9	5.88E-11

Table 4: **Psychological accuracy, subtractive test** – $\chi^2(1)$ and p values achieved by performing nested model analysis between the models *full-X* and the *full* model.

Additionally, we examine the coefficients of the surprisal predictions of each model. We extracted coefficients for individual surprisal measures independently from each of the models *base+X*. As shown in the columns *Indep* in Table 5, all coefficients are positive, implying, sensibly, that the more surprising a word, the longer time it takes to process that word.

Moreover, when all surprisal measures appear together in the same *full* model (columns *Joint*), we observe a consistent trend that the coefficients for DMPCFG and FG are positive, whereas that of the MAG is negative.

5.4 Discussion

Our results above indicate that the inference-based model provides the best account of our test data, both in terms of the likelihood it assigns to the test corpora and in terms of its ability to explain human fixation times. With respect to the full-parsing model this result is unsurprising. It is widely known that the conditional independence assumptions of PCFGs make them poor models of syntactic structure, and thus—presumably—of human sentence processing. Other recent work has shown that reasonable (though not state-of-the-art) parsing results can be achieved using models which relax the conditional independence assumptions of PCFGs by employing inventories of stored tree-fragments (i.e., *tree-substitution grammars*) similar to the fragment grammars model (De Marcken, 1996; Bod et al., 2003; Cohn et al., 2010; Post and Gildea, 2013; O'Donnell, 2015).

Models	Dundee		MIT	
	Indep.	Joint	Indep.	Joint
DMPCFG	5.94E-03	1.95E-03	8.08E-03	3.24E-03
MAG	1.00E-03	-1.41E-03	1.54E-04	-2.82E-03
FG	5.13E-03	5.49E-03	5.88E-03	6.97E-03

Table 5: **Mixed-effects coefficients** – the *Indep.* columns refer to the coefficients learned by the mixed-effects models *base+X* (one surprisal measure per model), whereas the *Joint* columns refer to coefficients of all surprisal measures within the *full* model.

The comparison with the full-listing model is more interesting. Not only does the full-listing model produce the worst performance of the three models in both corpora and for both evaluations, it actually produces negative correlations with reading times. We believe this result is indicative of a simple fact: while it has become clear that there is lexical storage of many syntactic constructions, and—in fact—the degree of storage may be considerably more than previously believed (Tremblay and Baayen, 2010; Bannard and Matthews, 2008)—syntax is still a domain which is mostly compositional. The full-listing model overfits, leading to nonsensical reading time predictions. In fact, this is likely a logical necessity—the vast combinatorial power implicit in natural language syntax means that even for a system with tremendous memory capacity, only a small fraction of potential structures can be stored.

6 Conclusion

In this paper, we have studied the ability of several models of computation and storage to explain

human sentence processing, demonstrating that a model which treates the problem as a case-by-case probabilistic inference provides the best fit to reading time datasets, when compared to two "limiting case" models which always compute or always store. However, as we emphasized in the introduction we see our contribution as primarily methodological. None of the models studied here represent state-of-the-art proposals for syntactic structure. Instead, we see these results together with the tools that we make available to the community, as providing a springboard for later research that will isolate exactly which factors, alone or in concert, best explain human sentence processing.

Acknowledgment

We gratefully acknowledge the help of Asaf Bachrach for making the MIT reading time dataset available to us. We thank Marten van Schijndel for discussion about mixed-effects models. Lastly, we thank the annonymous reviewers for their valuable comments and feedbacks.

References

R. Harald Baayen, Doug J. Davidson, and Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Asaf Bachrach, Brian Roark, Alex Marantz, Susan Whitfield-Gabrieli, Carlos Cardenas, , and John D.E. Gabrieli. 2009. Incremental prediction in naturalistic language processing: An fmri study.

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179 –190.

Colin Bannard and Danielle Matthews. 2008. Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science*, 19(3):241–248.

Douglas Bates, Martin Maechler, and Ben Bolker, 2011. *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-42.

Rens Bod, Remko Scha, and Khalil Sima'an, editors. 2003. *Data-Oriented Parsing*. CSLI, Stanford, CA.

Marisa F. Boston, John T. Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, 2(1):1–12.

Trevor Cohn, Phil Blunsom, and Sharon Goldwater. 2010. Inducing tree-substitution grammars. *Journal of Machine Learning Research*, 11:3053–3096.

Carl De Marcken. 1996. *Unsupervised Language Acquisition*. Ph.D. thesis, Massachusetts Institute of Technology.

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

Anna Maria Di Sciullo and Edwin Williams. 1987. *On the Definition of Word*. MIT Press, Cambridge, MA.

Jay Earley. 1968. *An Efficient Context-Free Parsing Algorithm*. Ph.D. thesis, Carnegie Mellon University.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538, September.

Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *CMCL*.

Stefan L Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–34.

Stefan L. Frank. 2009. Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *CogSci*.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *NAACL*.

Jennifer Hay. 2003. *Causes and Consequences of Word Structure*. Routledge, New York, NY.

Ray Jackendoff. 2002a. *Foundations of Language*. Oxford University Press, New York.

Ray Jackendoff. 2002b. What's in the lexicon? In S. Nooteboom, F. Weerman, and F. Wijnen, editors, *Storage and Computation in the Language Faculty*. Kluwer Academic Press, Dordrecht.

Frederick Jelinek and John D. Lafferty. 1991. Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17(3):315–323.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. In *NIPS*.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *NAACL*.

Alan Kennedy and Joel Pynte. 2005. Parafoveal-on-foveal effects in normal reading. *Vision Research*, 45(2):153–168.

Kenichi Kurihara and Taisuke Sato. 2006. Variational bayesian grammar induction for natural language. In *ICGI*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(2):1126–1177.

Minh-Thang Luong, Michael C. Frank, and Mark Johnson. 2013. Parsing entire discourses as very long strings: Capturing topic continuity in grounded language learning. *TACL*, 1(3):315–323.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn treebank: Annotating predicate argument structure. In *HLT*.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and semantic factors in processing difficulty: an integrated measure. In *ACL*.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Timothy J. O'Donnell, Noah D. Goodman, and Joshua B. Tenenbaum. 2009. Fragment grammars: Exploring computation and reuse in language. Technical Report MIT-CSAIL-TR-2009-013, MIT Computer Science and Artificial Intelligence Laboratory Technical Report Series, Cambridge, MA.

Timothy J. O'Donnell, Jesse Snedeker, Joshua B. Tenenbaum, and Noah D. Goodman. 2011. Productivity and reuse in language. In *CogSci*.

Timothy J. O'Donnell. 2015. *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. The MIT Press, Cambridge, Massachusetts.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *HLT-NAACL*.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *COLING-ACL*.

Matt Post and Daniel Gildea. 2013. Bayesian tree substitution grammars as a usage-based approach. *Language and Speech*, 56(3):291–308.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *EMNLP*.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27:249–276.

Andreas Stolcke. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, 21(2):165–201.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *ICSLP*.

Antoine Tremblay and R. Harald Baayen. 2010. Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In *Perspectives on Formulaic Language: Acquisition and Communication*, pages 151–173. Continuum.

Marten van Schijndel and William Schuler. 2013. An analysis of frequency- and memory-based processing costs. In *NAACL-HLT*.

Stephen Wu, Asaf Bachrach, Carlos Cardenas, and William Schuler. 2010. Complexity metrics in an incremental right-corner parser. In *ACL*.