The Influence of Syntactic Frequencies on Human Sentence Processing

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Marten van Schijndel, B.A.

Graduate Program in Linguistics

The Ohio State University

2017

Dissertation Committee:

William Schuler, Advisor

Micha Elsner

Shari Speer

Shravan Vasishth

Abstract

Humans are sensitive to the frequency of events, and this sensitivity is reflected in a wide range of behavioral and neural measures. This thesis focuses on the ways in which syntactic co-occurrence frequencies affect human language comprehension.

Previous psycholinguistic findings seemed to show that humans are not sensitive to verbal subcategorization frequencies. Instead, this work demonstrates that sensitivity to fine-grained syntactic frequencies provide a confounding explanation for those findings. A left-corner parser is defined that can be used to compute a variety of psycholinguistic complexity metrics in order to better control for such syntactic influences in future studies.

One of the strongest and most commonly used psycholinguistic measures output by the parser is surprisal (Hale, 2001; Levy, 2008), which estimates frequency-based comprehension difficulty based on the probability of an observation conditioned on the observations that preceded it. When used to predict reading times, however, this work shows that surprisal is mathematically inconsistent since it conditions on the immediately adjacent lexical material despite the fact that reading proceeds via saccades over non-adjacent material. This mathematical problem with surprisal can be corrected by summing surprisal over each saccade region to enable the measure to account for the probability of each new span of text conditioned on the preceding material that was actually observed. The corrected version of lexical (n-gram)

surprisal, cumulative n-gram surprisal, obtains a better fit to reading times than the uncorrected version, though the correction does not work for surprisal over syntactic (probabilistic context-free; PCFG) structure.

In addition to the frequency of observed events, this work explores the influence of frequency in how humans predict upcoming events. In particular, uncertainty about upcoming material (entropy) is shown to influence reading times, corroborating previous results in the literature (Roark et al., 2009; Angele et al., 2015). Unfortunately, the entropy over upcoming material is very expensive to compute, and so can be difficult to control for in psycholinguistic experiments. This work shows that the surprisal (n-gram and PCFG) of upcoming words, which is inexpensive to compute, can approximate the influence of that uncertainty on self-paced reading times.

The results in this thesis indicate that humans are sensitive to both lexical sequence frequencies and syntactic frequencies, and this work concludes by providing a proof-of-concept model of syntactic acquisition that links the two types of frequencies. The acquisition model demonstrates how a learner that is sensitive to linear ordering frequencies could end up acquiring long-distance dependencies, typically conceived as a hallmark of hierarchical syntax, in a fashion that replicates the acquisition timeline of children.

Acknowledgments

While a simple acknowledgement isn't much repayment for the huge amount of aid these people provided, hopefully it will help signal how much I appreciated everyone's help and friendship.

My advisor, William Schuler, was wonderful in far too many ways to describe here, but he was a great advisor, I have really enjoyed working with him, and I hope to someday approximate his incredible patience. I would also like to thank Micha Elsner for his mentorship and friendship. I really enjoyed collaborating with him on the work in Chapter 6, and hope to return to it at some point in the future.

Special thanks to Vera Demberg and her lab for taking a chance on my foray into neuroimaging and for hosting me in Saarbrücken. Despite the work not ultimately panning out, their feedback, especially that of Vera and Fatemeh Asr were instrumental in forming the direction this thesis ultimately went in. Shravan Vasishth and his lab have provided valuable feedback throughout this process, and Shravan's critical eye really helped sharpen up the argumentation and direction of this work.

Thanks to Stephanie Antetomaso, Evan Jaffe, Keeta Jones, and Noah Diewald, I was able to remain relatively human during this process, largely thanks to our collective insanity. I really appreciated their friendship, conversation, cooking, and humor and look forward to further adventures together in the future.

Best of luck to Manjuan Duan and her family who provided support and inspiration from the very beginning when we started the PhD program together. Manjuan is one of the most brilliant people I know and was a huge inspiration throughout.

Finally, thanks to Darcy for suffering through constant discussion of computational linguistics and for being a great partner and a wonderful mother. Her patience and compassion have really helped me keep perspective throughout this process and set a high bar to aspire to as a well-rounded human. And thanks to my ever-curious goblin-monkey, Viggo, who helped keep me grounded and continually reminded me of the joy of exploration and curiosity. Between the two of them, they have been helping inspire, amuse, and sustain me for roughly half my life, and I am eternally grateful. Sincerest thanks and love to both of them.

Thanks for helpful suggestions and feedback

Thanks to Matthew Traxler, Shari Speer, and the attendees of the Ohio State University syntactic processing seminar for feedback on the original idea for Chapter 2. Thanks to John Hale, David Reitter, and Jann Messer for helpful feedback on the work in Chapter 3. Thanks to Fatemeh Asr, Stefan Frank, Nathaniel Smith, Vera Demberg, Rick Lewis, and Roger Levy for helpful comments and suggestions regarding the work in Chapter 4. Thanks to Mike White, Micha Elsner, and Tal Linzen for useful commentary and feedback on the work in Chapter 5. Thanks to Peter Culicover, William Schuler, Laura Wagner, and the attendees of the OSU 2013 Fall Linguistics Colloquium Fest for feedback on the work in Chapter 6.

Thanks to the funders of this work

The work in Chapter 3 was funded by an Ohio State University Graduate Research Fellowship. The work in Chapters 2 and 6 was funded by an Ohio State University Department of Linguistics Targeted Investment for Excellence (TIE) grant for collaborative interdisciplinary projects conducted during the academic year 2012-13. The work in Chapters 4 and 5 was supported by a National Science Foundation Graduate Research Fellowship Program Award DGE-1343012.

Vita

2009 ..... B.A. Linguistics (Cum Laude),
Western Washington University.

2017 ..... Ph.D. Linguistics,
The Ohio State University.

Publications

Research Publications

Marten van Schijndel and William Schuler "Approximations of predictive entropy correlate with reading times". Proceedings of CogSci 2017. 2017.

Rajakrishnan Rajkumar, Marten van Schijndel, Michael White, and William Schuler "Investigating Locality Effects and Surprisal in Written English Syntactic Choice Phenomena". Cognition, 155: 204–232. 2016.

Alessandra Zarcone, Marten van Schijndel, Jorrig Vogels, and Vera Demberg "Salience and attention in surprisal-based accounts of language processing". Frontiers in Psychology, 7 (844). 2016.

Marten van Schijndel and William Schuler "Addressing surprisal deficiencies in reading time models". Proceedings of the Computational Linguistics for Linguistic Complexity Workshop (CL4LC 2016). 2016.

Cory Shain, Marten van Schijndel, Edward Gibson, and William Schuler "Memory access during sentence processing causes reading time latency". Proceedings of the Computational Linguistics for Linguistic Complexity Workshop (CL4LC 2016). 2016.

Marten van Schijndel and William Schuler "Hierarchic syntax improves reading time prediction". Proceedings of NAACL 2015. 2015.

Marten van Schijndel, Brian Murphy, and William Schuler "Evidence of syntactic working memory usage in MEG data". Proceedings of CMCL 2015. 2015.

Evan Jaffe, Lifeng Jin, David King, and Marten van Schijndel "Azmat: Sentence similarity using associative matrices". Proceedings of the International Workshop on Semantic Evaluation (SemEval 2015). 2015.

Marten van Schijndel and Micha Elsner "Bootstrapping into filler-gap: An acquisition story". Proceedings of ACL 2014. 2014.

Marten van Schijndel, William Schuler, and Peter Culicover "Frequency Effects in the Processing of Unbounded Dependencies". Proceedings of CogSci 2014. 2014.

Marten van Schijndel and William Schuler "An Analysis of Frequency- and Memory-Based Processing Costs". Proceedings of NAACL 2013. 2013.

Marten van Schijndel, Andy Exley, and William Schuler "A Model of Language Processing as Hierarchic Sequential Prediction". Topics in Cognitive Science, 5 (3): 522–540. 2013.

Marten van Schijndel, Luan Nguyen, and William Schuler "An Analysis of Memory-based Processing Costs using Incremental Deep Syntactic Dependency Parsing". Proceedings of the 4th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2013). 2013.

Luan Nguyen, Marten van Schijndel, and William Schuler "Accurate Unbounded Dependency Recovery using Generalized Categorial Grammars". Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). 2012. Received 'Best Paper' at COLING 2012

Marten van Schijndel, Andy Exley, and William Schuler "Connectionist-Inspired Incremental PCFG Parsing". Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012). 2012.

Fields of Study

Major Field: Linguistics

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Previous studies have questioned the degree to which hierarchical syntax is used by humans during online sentence processing (Frank and Bod, 2011; Frank et al., 2012). However, in order to know whether a given result can be attributed to an influence of latent structural syntactic frequencies, it is first critical to adequately control for the observed lexical sequence frequencies. This thesis provides an in-depth investigation of lexical frequency effects (n-grams), syntactic frequency effects, and the degree to which syntactic frequencies predict behavioral data once the n-gram frequencies have been controlled for.

Chapter 2 demonstrates that hierarchical syntax provides a compelling potential confound to many previous findings, suggesting a need to control for hierarchic occurrence frequencies in psycholinguistic studies. Chapter 3 describes a left-corner parser that can be used to compute a variety of incremental complexity measures, including occurrence frequency estimates of hierarchical syntactic structures. Chapter 4 identifies and corrects an inconsistency in how surprisal (both n-gram and probabilistic context-free grammar; PCFG) is typically used in reading time studies to estimate incremental processing complexity due to frequency effects. Correcting this inconsistency produces a baseline n-gram model that better captures frequency influences on reading times. Once n-gram surprisal is improved, PCFG surprisal becomes a much

weaker predictor of reading times, though at least some of its remaining predictivity seems due to long-distance dependencies, which would be much harder to capture with n-grams. Chapter 5 shows that humans are sensitive to upcoming (un)certainty and that the degree of certainty about future observations (both lexical and syntactic) affects self-paced reading times. The chapter further demonstrates that a single sample from the conditional probability distribution over future observations (future surprisal) may be used as an inexpensive aggregate entropy approximation that provides more psycholinguistically robust entropy estimates than the typical point-wise entropy estimates which estimate uncertainty at each time-step by hallucinating every possible continuation. Finally, Chapter 6 describes an acquisition model that accounts for the acquisition pattern of filler-gap dependencies, providing a framework for bootstrapping from a linear frequency sensitivity into a processing model that could be sensitive to hierarchical syntactic frequencies. The results in this thesis indicate that, though PCFG surprisal is only weakly predictive of first-pass reading times after theoretically-motivated n-gram adjustments, hierarchical syntactic frequencies do correlate with self-paced reading times even over strong lexical sequence predictors, suggesting that humans are sensitive to hierarchical syntactic frequencies during processing.

Chapter 2: The necessity of hierarchical syntax as a frequency control*

Psycholinguistic studies are subject to strong frequency influences, so psycholinguists often must either statistically or experimentally control for possible frequency effects confounded with the effects of interest. Statistical frequency controls provide explicit baseline terms that quantify the expected frequency influence. Systematic deviations from the baseline predictions indicate some kind of processing effect divorced from the frequency effects linguists already know about. Experimental frequency control occurs prior to data collection by tightly controlling the stimuli that are used in the study. For example, stimuli may be generated so that the lexical context preceding a target region is present across all experimental conditions in order to remove the possibility that an observed effect will be caused by the different preceding contexts rather than by the target manipulation. This chapter demonstrates that hierarchical syntactic frequency is a strong potential confound for a variety of previous studies. These results suggest that cloze norming should generally be augmented by corpus statistics to provide adequate frequency controls.

*The work in this chapter originally appeared in van Schijndel et al. (2014).

2.1 Introduction

Unbounded dependencies (e.g., between the book and about in the noun phrase [the book]$_i$ the author wrote about t_i) consist of a filler (the book) and an attachment site or gap (t_i) which can be separated by an unbounded number of words. Since gaps are not overtly represented in sentences, their locations can be temporarily ambiguous (e.g., after wrote or after about). Some researchers have suggested that maintaining such dependencies introduces additional processing difficulty (Chomsky and Miller, 1963; Gibson, 2000). In order to quickly resolve ambiguous unbounded dependencies and ease any potential difficulty, one might expect readers to make full use of the information at their disposal to complete unbounded dependencies as soon as possible.

Several self-paced reading and eye-tracking studies have explored whether readers make use of subcategorization preferences of verbs in order to immediately restrict the hypothesis space of unbounded dependency attachments (Mitchell, 1987; Pickering et al., 2000; van Gompel and Pickering, 2001; Pickering and Traxler, 2003). Subcategorization preference or bias may be ascertained by observing how frequently a verb appears with given argument types (a verb that appears very frequently with a noun phrase (NP) direct object but occasionally without any direct object argument would be deemed an optionally transitive verb with a transitive bias to take NP arguments). The following sentences from Pickering and Traxler (2003) are representative of the stimuli used in such experiments:

(1) That's the plane that the pilot landed carefully behind in the fog at the airport.

(2) That's the truck that the pilot landed carefully behind in the fog at the airport.

These authors claim that if readers use subcategorization frequency in processing, the implausibility in (2) of truck as an argument of landed should not cause readers to slow since landed prefers a prepositional phrase (PP) argument to a noun phrase (NP) argument. Instead, readers of (2) do slow down compared to readers of (1) after reading landed, which the authors claim suggests that they have difficulty with the implausible interpretation of (2) that arises from the attachment of the unbounded dependency to the verb in spite of its subcategorization bias (pilots don't usually land trucks). These previous studies have interpreted such results as an indication that subcategorization frequency is not used by readers when initially resolving unbounded dependencies; rather, readers seem to employ a simple early-attachment heuristic.

This chapter reviews recent articles from the psycholinguistics literature which suggest an alternative, frequency-based explanation for this finding. It then goes on to show how a probabilistic context free grammar (PCFG) can be constructed from corpora annotated with unbounded dependencies and used to estimate the frequency effects involved in unbounded dependency processing. This analysis shows that the slow-down in sentences such as (2) may be explained by the frequencies of non-preterminal syntactic configurations which may have a much stronger impact than subcategorization preferences.

2.2 Background

Since Pickering and Traxler (2003), a number of studies have revisited the claim that subcategorization is not used by readers in initial attachment of unbounded

dependencies. For example, Staub et al. (2006) conducted two experiments that explored the time course of processing a particular kind of unbounded dependency: heavy-NP shift constructions. They found that sentences containing an optionally transitive verb with a transitive bias (e.g., The teacher helped immediately [the confused student]) were processed more slowly upon encountering the shifted region than sentences containing an obligatorily transitive verb (e.g., The teacher corrected immediately [the unusual answer]). They interpret their results as evidence that readers adopt a parsing heuristic that disprefers a heavy-NP shift interpretation rather than purely relying on the subcategorization bias of the verb.[2] Otherwise, verbs with a transitive bias would force an initial transitive reading to be adopted and so would not yield this pattern of slowing. They point out, however, that their results may have been driven by the infrequency of heavy-NP shift as a construction. That is, the infrequency of heavy-NP shift may overwhelm any transitivity bias of the verb. The PCFG analysis described in this chapter may be construed as a formalization of this analysis.

Arai and Keller (2013) suggest a similar frequency explanation of the findings of Pickering and Traxler (2003) based on visual world experiments similar to those of Altmann and Kamide (1999), Kamide et al. (2003a), and Kamide et al. (2003b). By observing where subjects' eyes moved as they listened to sentences containing transitive or intransitive verbs, they found evidence that subjects do take selectional information into account at the verb. Specifically, plausible arguments and complements are fixated on more frequently than implausible ones when verbs with either subcategorization bias are heard. Based on this finding, they speculated that the findings of

[2]While Staub et al. (2006) argue for a serial model of language processing, this chapter remains agnostic with respect to whether processing is serial or parallel.

studies such as Pickering and Traxler (2003) could be due to the frequency of main clause direct object constructions when compared with the alternative constructions supported by verbal subcategorization, but they did not evaluate this claim.

Finally, Staub (2007) conducted three self-paced reading studies in an attempt to remove confounds from Pickering et al. (2000) and similar studies. By separating the intransitive verbs used in those studies into unaccusative verbs (e.g., erupt), which can never take a direct object argument, and unergative verbs (e.g., sneeze), which can take direct object arguments under particular circumstances, Staub was able to construct a set of sentences that were truly obligatorily intransitive. When reading a sentence containing an obligatorily intransitive (unaccusative) verb, readers did not show any evidence of attaching the filler of the unbounded dependency to the verb, unlike in the unergative case where there was a slight chance of obtaining a transitive interpretation. This finding indicates that any possibility of transitive interpretation, even when that possibility is very slight, can cause readers to adopt implausible analyses, which suggests that the frequency of a direct object interpretation can overwhelm the lexically-specific bias of a verb.

This chapter directly investigates these recent claims that earlier findings of insensitivity to verb subcategorization bias may be due to syntactic configuration frequency. If the probability of a syntactic configuration is defined as the product of the probabilities of its component syntactic configurations and its lexical items, a very small or very large syntactic probability (e.g., that of heavy-NP shift, or the prevalence of direct object complements) could overwhelm verb-specific argument biases.

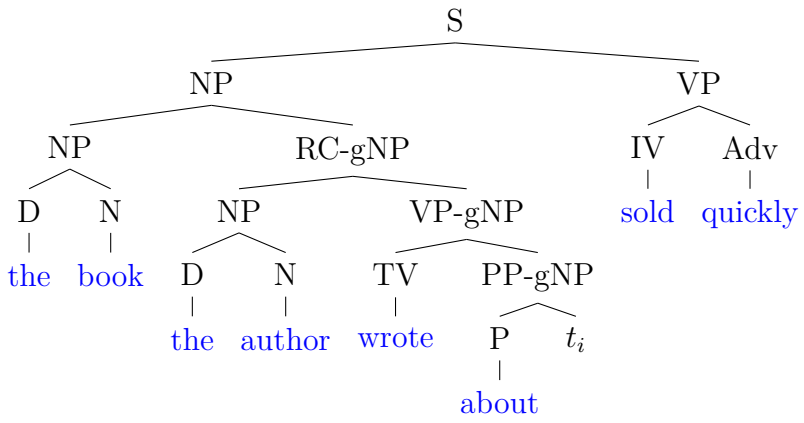

Figure 2.1: An example syntactic analysis of The book the author wrote about sold quickly with a GCG-like treatment of unbounded dependencies. The gap is annotated with t_i in the figure, only. Each category is sensitive to whether it has an unresolved gap within its subtree.

2.3 Probabilistic Grammars

Probabilities for syntactic configurations can be obtained by assigning probabilities to grammar rules. For example, a prepositional phrase (PP) usually generates a preposition (P) and a noun phrase (NP). Each such rule in the grammar may be assigned a conditional probability based on the frequency with which that parent category generates those child categories in large corpora. The resulting probability-weighted grammar is called a probabilistic context-free grammar (PCFG, Booth & Thompson, 1973). The probability of a syntactic configuration can then be estimated as the product of these conditional rule probabilities.

Well-studied algorithms exist for finding and refining PCFGs from data (Petrov et al., 2006), and PCFGs have been shown to be useful as a basis for information-theoretic accounts of garden path effects and reading time delays (Jurafsky, 1996; Hale, 2001, 2003, 2006; Levy, 2008). Usually, however, PCFG models have excluded

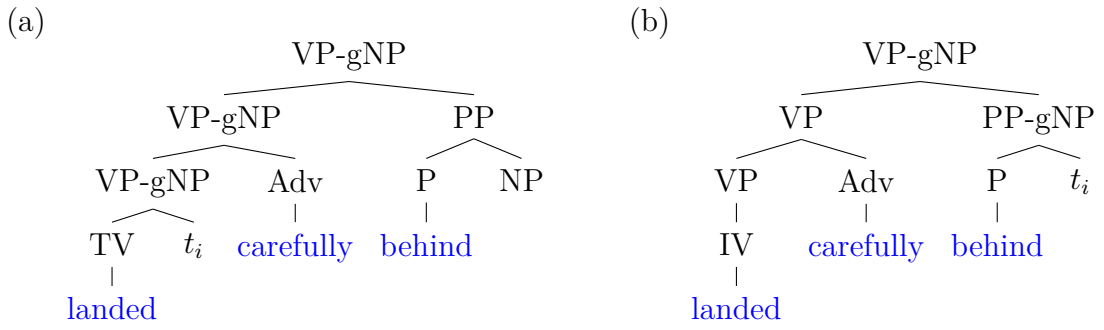

Figure 2.2: Alternative parses of a portion of That's the plane/truck that the pilot landed carefully behind in the fog at the airport, shown immediately after observing the word behind. The predicted syntactic category of the next observation is shown, and gaps are annotated with t_i. Parse (a) corresponds to a transitive NP interpretation and Parse (b) corresponds to an intransitive PP interpretation.

unbounded dependency information because of its inherent complexity. In order to capture unbounded dependency information and still use existing algorithms for obtaining highly accurate PCFGs, this chapter uses a generalized categorial grammar (GCG) (Bach, 1981; Nguyen et al., 2012), which passes unbounded dependencies from parents to children and so makes the propagation of a gap into a category context-free (solely dependent on whether a gap exists in the parent category and on whether the preceding sibling could serve as a filler).

The Nguyen et al. (2012) GCG encodes gap information using a -g operator added to categories that contain a gap (see Figure 2.1), so a verb phrase (VP) with a gapped NP argument would be assigned the category VP-gNP and would expand to a child transitive verb (TV) and a gap associated with an NP. To link this gap to the correct filler, this GCG propagates the -g from the sibling category of the filler to each appropriate child in the syntax tree in a fashion similar to the SLASH category of

Interpretation	Grammar rule	Prob
Transitive	VP-gNP→VP-gNP PP	0.17
Intransitive	VP-gNP→VP PP-gNP	0.01

Table 2.1: The probability of the grammar rules associated with transitive and intransitive interpretations during incremental resolution of unbounded dependencies as calculated from the Wall Street Journal text corpus. These numbers are based on the 2,355 occurrences of VP-gNP in the corpus.

Head-driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994) and other HPSG-like context-free gap notations (Hale, 2001; Lewis and Vasishth, 2005).

2.4 Evaluation

Aside from verb subcategorization bias, the difference between the transitive and intransitive interpretations of the sentences in Pickering and Traxler (2003) and related studies is that a transitive interpretation hypothesizes the gap as the complement of the main verb, whereas an intransitive interpretation hypothesizes the gap as the complement of the preposition (see Figure 2.2). In order to quantify the frequency interactions that may be behind the findings of studies such as Pickering and Traxler (2003), the Wall Street Journal (WSJ) portion of the Penn Treebank corpus of English (Marcus et al., 1993) is reannotated using a GCG as described by Nguyen et al. (2012).

Counts of each syntactic configuration in this reannotated corpus indicate that the intransitive interpretation is much less frequent than the transitive interpretation (see Table 2.1). Since the parse of each interpretation is otherwise equivalent up to the verb, the probability of subjects entertaining each possible interpretation may

be computed by taking the product of each grammar rule probability in Table 2.1 and the probability of a subsequent rule generating the verb from a given preterminal category (since all other relevant grammar rules are common to both interpretations). The probability of a verb being generated from a given preterminal category (TV or IV) is proportional to ($\propto$) the subcategorization bias of the verb divided by the prior probability of the preterminal:[3]

$$\begin{aligned} \mathrm{P}(\text{Trans}) &= \mathrm{P}(\text{VP-gNP} \rightarrow \text{VP-gNP} \quad \text{PP}) \cdot \mathrm{P}(\textit{verb} \mid \text{TV}) \\ &\propto \mathrm{P}(\text{VP-gNP} \rightarrow \text{VP-gNP} \quad \text{PP}) \cdot \frac{\mathrm{P}(\text{TV} \mid \textit{verb})}{\mathrm{P}(\text{TV})} \qquad (2.1) \\ \mathrm{P}(\text{Intrans}) &= \mathrm{P}(\text{VP-gNP} \rightarrow \text{VP} \quad \text{PP-gNP}) \cdot \mathrm{P}(\textit{verb} \mid \text{IV}) \\ &\propto \mathrm{P}(\text{VP-gNP} \rightarrow \text{VP} \quad \text{PP-gNP}) \cdot \frac{\mathrm{P}(\text{IV} \mid \textit{verb})}{\mathrm{P}(\text{IV})} \qquad (2.2) \end{aligned}$$

Counts from the reannotated corpus show that 14719 transitive verbs (TV) were used and 5617 intransitive verbs (IV) were used, which gives relative prior probabilities for each type of verb that can be used as normalizing constants in the evaluation:

$$\mathrm{P}(\text{TV}) = \frac{\text{count}(\text{TV})}{\text{count}(\text{IV}) + \text{count}(\text{TV})} = \frac{14719}{5617 + 14719} = 0.72 \qquad (2.3)$$

$$\mathrm{P}(\text{IV}) = \frac{\text{count}(\text{IV})}{\text{count}(\text{IV}) + \text{count}(\text{TV})} = \frac{5617}{5617 + 14719} = 0.28 \qquad (2.4)$$

For ease of comparison, this chapter makes use of the bias frequencies obtained during the verb norming study of Pickering and Traxler (2003), which were obtained

[3]An additional adverb (e.g., carefully) is sometimes used to increase the ambiguous region of each sentence during reading experiments. Although the probabilities for rules with and without the adverb are different, including the probabilities of adverbial rules (VP $\rightarrow$ VP Adv and VP-gNP $\rightarrow$ VP-gNP Adv) and the probabilities of preterminal rules (VP $\rightarrow$ IV and VP-gNP $\rightarrow$ TV) does not change the direction of the effect reported in this chapter and generally increases the magnitude (with preterminal rules, the probability of transitive interpretation: 0.046 and intransitive interpretation: 0.0001; with adverbial rules, probability of transitive interpretation: 0.0078 and intransitive interpretation: 0.0001), so they are omitted for clarity.

by asking 90 subjects to write sentences containing the relevant verbs and counting the number of times a verb appeared in a given configuration.[4] As an example, landed appeared with an NP 25% of the time, a PP 40% of the time, and with neither 35% of the time. Using the above formula of rule·bias/prior, this means the probability of landed inducing a transitive NP complement interpretation in subjects is $0.17 \cdot 0.25/0.72 = 0.059$ compared with the probability of landed inducing the intransitive PP complement interpretation in subjects, which is $0.01 \cdot 0.4/0.28 = 0.014$. The NP complement interpretation of landed is thus 400% more likely for subjects to adopt than a PP complement interpretation, despite the prima facie bias for landed to take a PP complement. This disparity directly arises from the substantially greater probability of propagating a gap dependency to a VP child than to a PP child. On average, the PP-biased verbs used in Pickering and Traxler (2003) have an intransitive bias of 0.52 and a transitive bias of 0.14, which means the average PP-biased verb is nearly twice as likely to induce a transitive interpretation than an intransitive interpretation in subjects (NP interpretation: $0.17 \cdot 0.14/0.72 = 0.033$; PP interpretation: $0.01 \cdot 0.52/0.28 = 0.019$). Even the second most PP-biased verb used by Pickering and Traxler (2003), searched, which appeared with an NP 15% of the time and with a PP 75% of the time, is more likely to receive an NP interpretation than a PP interpretation (NP interpretation: $0.17 \cdot 0.15/0.72 = 0.035$; PP interpretation: $0.01 \cdot 0.75/0.28 = 0.027$). Lacking a representative number of verbs with a strong enough subcategorization bias to induce a PP-interpretation, it is unsurprising that such studies have failed to observe an effect of verb subcategorization bias.

[4]Pickering and Traxler (2003) also determined the subcategorization bias of each verb using other norming studies, but the study that yielded the results used in this chapter had the largest subject pool. Using one of their other sets of bias results does not significantly affect the results of this chapter.

2.5 Discussion

A possible criticism of using frequency probabilities derived from the WSJ corpus is that the lexicon or the distribution of syntactic configurations may not generalize well to other domains (Sekine, 1997; McClosky, 2010). However, the lexeme-specific probabilities used in this study were determined experimentally by Pickering and Traxler (2003), so they do not depend on the WSJ lexicon. Only the syntactic rule probabilities are derived from the WSJ corpus; however, Nguyen et al. (2012) showed that a parser trained only on a GCG-reannotated WSJ corpus can achieve state-of-the-art parsing accuracy for unbounded dependencies in a variety of domains (news, narrative, etc). This finding suggests the distribution of unbounded dependencies in the WSJ corpus is representative of the distribution of English unbounded dependencies as a whole.

The same probability model given in this chapter can be used to account for the findings of Staub (2007) that readers do not mistakenly attach fillers to unaccusative verbs (e.g., erupt). Since unaccusative verbs cannot take an NP argument, the probability of erupt inducing an NP transitive interpretation is $0.17 \cdot 0.0/0.72 = 0.0$.

Further, this model can account for the findings of Staub et al. (2006) regarding reading times of heavy-NP shift constructions. Though it is beyond the scope of this chapter to detail the syntactic analyses that are involved, heavy-NP shift constructions are less frequent than unshifted constructions. Using a similar analysis to that given here, it may be shown that this model replicates the findings of Staub et al. (2006): obligatorily transitive verbs should cause readers to slow at the inserted material in shifted constructions (because shifted constructions are less frequent than unshifted constructions) and optionally transitive verbs should cause readers to slow

at the shifted NP (because the infrequency of shifted constructions outweighs all but the strongest transitive biases). Interestingly, preliminary results exploring heavy-NP shift with this model indicate that optionally transitive verbs with a transitive bias of around 87% may yield a slow-down at the inserted adverb (when compared with adverbs in unshifted, optionally transitive constructions) rather than the object noun since that optional transitive bias should outweigh the bias of intransitive constructions but not completely outweigh the preference to not shift. Such a finding was not observed by Staub et al. (2006), but their optionally transitive verbs did not approach this level of transitive bias.[5]

The effectiveness of this model at accounting for a variety of experimental findings has potential implications for theories of human sentence processing. For example, this model assumes that subcategorization information (e.g., the number and type of required arguments) is present immediately during parsing, regardless of its regularity. In contrast, some theories of sentence processing like the Garden Path Model (Frazier, 1987) or Construal (Frazier and Clifton, 1996) posit that only regular grammatical patterns (e.g., transitive verbs) are immediately available to the parser, whereas irregular exceptions only become available during a later stage in processing. Such theories have typically been supported by findings (e.g., Pickering & Traxler, 2003, and Pickering et al., 2000) that subcategorization is not immediately used during sentence processing. While the present study does not rule out multi-stage processing models altogether, it does show that processing can make immediate use of subcategorization

[5]A script to replicate all findings in this chapter is available at https://github.com/vansky/unbounded_frequency. The replication script also confirms the preliminary heavy-NP shift analysis given here.

biases and still replicate findings which had been interpreted as showing that subcategorization is not used immediately during processing. Therefore, a pool of supporting evidence that was previously thought to strongly favor multi-stage processing models should no longer be considered to do so.

Conclusion

While it may be true that verbs have specific subcategorization preferences, this chapter has shown that the overwhelming bias to propagate a gap into a verb phrase rather than a prepositional phrase sibling will override all but the strongest subcategorization preferences during online processing. In fact, an optionally transitive verb would have to appear with a PP 6.6 times for every 1 NP (85% intransitive bias) in order to have an even chance of inducing a PP complement interpretation compared with an NP complement interpretation. This work, therefore, provides quantitative evidence in support of recent suggestions (Staub et al., 2006; Staub, 2007; Arai and Keller, 2013) that previous findings of reader insensitivity to verb subcategorization preference may be due to the frequencies of the syntactic configurations involved. Further, these results demonstrate that cloze probability norming studies only account for a single term in the larger equation that describes the incremental expectations of readers. This finding highlights the need to account for frequency at multiple levels of processing rather than simply in terms of lexical biases.

Chapter 3: Computing incremental complexity measures*

In order to control for syntactic frequency effects on broad-coverage data, this chapter describes an incremental parser that computes a variety of word-level complexity measures. Psycholinguistic studies can use the parser either to control for possible frequency confounds, as outlined in the previous chapter, or to explicitly look for correlations between those complexity estimates and behavioral measures such as reading times.

3.1 Incremental parsing

Whether visually, auditorily, or hapticly, language is a sequential process, and humans process language incrementally as input is received. Evidence for the incremental processing of language comes from garden path sentences (e.g., Frazier and Rayner, 1982; Slattery et al., 2013), eye-tracking studies (e.g., Altmann and Kamide, 1999; Ito and Speer, 2008; Arai and Keller, 2013), and brain imaging studies (e.g., Kutas and Hillyard, 1980; Frank et al., 2015) as well as a variety of other sources.

One of the most intuitive ways to model human sentence processing, then, is as an incremental parsing operation. The present work will largely restrict its scope

*The work in this chapter originally appeared in van Schijndel et al. (2013a).

to syntactic processing, which covers how the underlying structural relationships between words are constructed. In particular, this chapter will describe a probabilistic incremental parser. Probilistic models can not only operate at scale, but they can also typically estimate incremental prefix probabilities for analyses. Prefix probabilities are needed to compute information-theoretic complexity measures, which have seen growing popularity among psycholinguists (Hale, 2001, 2006; Levy, 2008; Demberg and Keller, 2008; Frank et al., 2015).

Recent models of working memory (Howard and Kahana, 2002; Botvinick, 2007) are defined in terms of hierarchic sequential and temporal cueing operations. Observed events (for example, visible grasping and manipulation actions) are organized into hypothesized sequences of more general states (actions in a process of making coffee), encoded in a changing context (a set of continuous-valued neural units expressing features of states) representing a weighted set of active hypotheses pursued in parallel. Sequences of these states may themselves belong to higher-level sequences (steps in making breakfast, for example), forming a multi-level hierarchy. Sequential transitions between states in each level of this hierarchy may be directly learned from experience, then recalled rapidly and reliably by 'sequentially' cueing successive states on features of preceding states (observations of pouring coffee into a mug are likely to be followed by adding milk, say). But when these learned sequences terminate, the process must recall its place in some immediately superordinate sequence (the current step in making breakfast). Unlike content-based sequential cueing, the transition from a terminating subordinate state to a state in some immediately superordinate sequence may not have been directly learned, so the superordinate state must instead be recalled based on the similarity of a set of temporal features associated with this

state to a set of temporal features in the current context. These temporal features change as the current context changes. As a result, this 'temporal cueing' becomes less reliable and takes more time to converge as the similarity of these temporal features decreases. This provides a tidy explanation of scale-invariant long-term recency effects in serial recall experiments (Bjork and Whitten, 1974; Crowder, 1982).[6]

Can sentence processing be modeled using the same kind of sequential and temporal cueing operations popular in the computational memory community? Words, phrases, and sentences form hierarchies just like observed events, actions, and processes. But unlike phrase structure trees and discourse structures, the hierarchic sequence models described in the computational memory literature are typically fairly shallow, reflecting the tendency for memories of superordinate sequence states to become increasingly conflated as the temporal features of the current context diverge from their temporal features.

This chapter describes a broad-coverage probabilistic sentence processing model that uses an optionally arc-eager variant of a left-corner parsing strategy (Rosencrantz and Lewis II, 1970; Johnson-Laird, 1983; Abney and Johnson, 1991; Schuler et al., 2010) to flatten sentence processing operations into a similarly shallow hierarchy of learned sequences. These sequences are then mapped to explicit states in a hierarchic probabilistic sequence model. Unlike the similar broad-coverage sequence model parsers of Crocker and Brants (2000) and Henderson (2004), this model exploits a property of optionally arc-eager left-corner parsing that ensures that subordinate sequences are initiated or terminated no more than once in each hypothesis after each

[6]That is, subjects show a preference for recalling recent items in list recall studies even when these items are separated by distractor tasks (e.g. performing arithmetic), contra predictions of working memory models that posit short-term buffers to explain recency effects.

observed word. This provides a natural constraint on temporal cueing operations, yielding a shallower hierarchy of sequence states than those required by Crocker and Brants (2000) and Henderson (2004).

The main result of this paper is that a broad-coverage model with these kinds of constraints can process large newspaper corpora with the same accuracy as a state-of-the-art parser not defined in terms of hierarchic sequential working memory operations. As argued by Crocker and Brants (2000), broad-coverage models like this are valuable because they allow experimental evaluation of interactions among factors across a variety of phenomena under uniform modeling assumptions, providing a more rigorous test of claims about general linguistic behavior, including unanticipated consequences of modeling decisions. This is expected to facilitate broad-coverage experimental evaluations in which memory-based measures (for example, measures of subordinate sequence termination as a proportion of the set of active hypotheses) can be combined with frequency-based measures (for example, surprisal or entropy reduction; Hale, 2001, 2006) on a fair footing.

The remainder of this chapter describes an incremental processing model based on sequential and temporal cueing operations, adapts the model for use with probabilistic context-free grammars, then describes a broad-coverage evaluation of this model in which the model performs approximately as well as a non-incremental baseline.

3.1.1 Model

The model described in this chapter is defined in terms of nested signs, which may consist of phrases (e.g. bought the rights), clauses (e.g. the studio bought the rights), and non-constituent signs (e.g. the studio bought) with various syntactic categories.

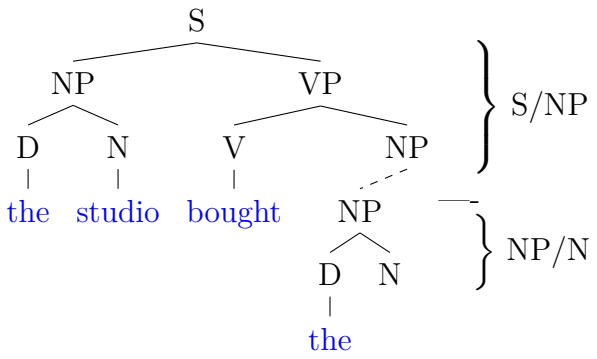

Figure 3.1: Two disjoint connected components of a syntactic tree structure for the sentence the studio bought the publisher's rights, shown immediately prior to the word publisher.

These signs can be arranged graphically in a binary tree structure, showing how they are nested.[7] This tree structure can be incrementally recognized with minimal memory requirements using an optionally arc-eager variant of a left-corner parsing process (Rosencrantz and Lewis II, 1970; Johnson-Laird, 1983; Abney and Johnson, 1991; Schuler et al., 2010), which stores at each incremental step a set of states delimiting connected components in this tree structure that cannot yet be directly merged. This process uses stored states only within center-embedded structures (if [$_S$ either [$_S$ …] or …] then …), not left- or right-embedded structures ([$_{NP}$ [$_{NP}$ the bride]'s mother]'s friend, or the cat that [$_S$ chased the rat that [$_S$ ate the malt]]). As a result, this memory hierarchy is much shallower than the tree representation of the syntactic relationships that are recognized.

The model described in this chapter represents time in discrete steps t, corresponding to discrete observations of words x_t. At each time step, the model maintains

[7]Non-binary trees may be binarized through insertion of unique intermediate categories.

several hypotheses q_t which are probabilistically weighted and considered in parallel, as an explicit decomposition of the high-dimensional hidden context of a recurrent neural network like that of Botvinick (2007). Each hypothesis defines a hierarchy of sequence states q_t^d, ordered by depth d from superordinate ($d = 1$) to subordinate ($d > 1$). Each sequence state q_t^d defines a maximal connected component of predicted syntactic tree structure a_t^d/b_t^d, consisting of an active sign of category a_t^d lacking an awaited sign of category b_t^d yet to come. Any connected sequence of signs descending over time in a predicted syntactic tree structure (e.g. S, VP, and NP in Figure 3.1) can form a single connected-component state (e.g. S/NP).

Over time, the model predicts syntactically-structured signs, compares them against observed words, generalizes them into new connected-component states, then merges them with subordinate and superordinate connected-component states as the syntactic relations that connect them in the syntactic tree structure are predicted. Any sequence of hierarchically-organized connected-component states generated by this model corresponds to a traversal of a predicted syntactic tree structure. For example, Figure 3.2 shows a hierarchic state sequence corresponding to a traversal of a predicted syntactic tree structure for the sentence The studio bought the publisher's rights.

Note that transitions within a single hierarchy level may predict syntactic relations upward along sequences of initial children (from the NP the publisher to the D the publisher's between time steps 5 and 6, for example), and downward along sequences of final children (from the VP bought the publisher's rights to the NP the publisher's rights between time steps 3 and 4). Although this predicted syntactic structure may have an unbounded number of recursive initial or final children (for example,

sequences of possessives extending the initial portion of a noun phrase or sequences of adjectives extending the final portion of a noun phrase, as shown in Figure 3.3), it requires only a bounded number of connected-component states at any given time step. This flattening of the syntactic tree structure into potentially cyclic sequences of connected-component states is similar to the programming strategy of replacing head recursion and tail recursion with loops (where program instructions correspond to states, recursive function calls in the call stack correspond to subordinate states in the state hierarchy, and loops correspond to cyclic transitions over states like S/N and D/G). This is also similar to the left-corner parsing strategy commonly used in sentence processing models (Johnson-Laird, 1983; Abney and Johnson, 1991; Gibson, 1991; Henderson, 2004; Lewis and Vasishth, 2005), except that connected components in the state hierarchy defined here are paired into active signs with awaited signs somewhere on their final (right) edge, rather than as awaited signs with active signs somewhere on their initial (left) edge.[8]

Sequence modeling with connected components

The general hierarchic sequence model described in this chapter is defined in terms of a set of syntactic relations — in particular, a set of context-free rules of the form $a \rightarrow a'\ b'$ (meaning sign a is composed of an initial child sign a' followed by final child sign b'), or $a \rightarrow x$ (meaning sign a is associated directly with an observation of word x). In addition to this set of syntactic relations, this model also assumes an ability to predict arbitrarily deeply nested initial descendant sub-signs a' of larger signs b, denoted $b \overset{+}{\rightarrow} a' \ldots$.

[8]In previous work, this is therefore referred to as right-corner parsing (Schuler et al., 2010).

A simple nondeterministic process for incrementally predicting phase structures using a sequence of connected-component states can be defined as a deductive system, given an input sequence consisting of a top-level connected-component state $\top/\top$, corresponding to an existing discourse context, followed by a sequence of observed words $x_1, \ldots, x_n$, processed in time order.[9] As each x_t is encountered, it is connected to the existing components, or it introduces a new disjoint component using productions that treat each word as the first observation of a newly initiated connected-component state, or as the last observation of a terminated connected-component state, or as neither, or as both.

First, if an observation x_t can attach as the awaited sign b of the most recent (most subordinate) connected component a/b from a to b, it is hypothesized to do so, turning this incomplete connected component into a complete connected component below a (via production F–, below); or if the observation can serve as a lower descendant of this awaited sign b, it is hypothesized to form the first sign a' in a newly initiated complete connected component (via production F+):

$$\frac{a/b \quad x_t}{a}\ b \rightarrow x_t \tag{F–}$$

$$\frac{a/b \quad x_t}{a/b \quad a'}\ b \xrightarrow{+} a' \ \ldots\ ; \quad a' \rightarrow x_t \tag{F+}$$

Then, if either of these complete connected components (below a or a' above, matched to a'' below) can attach as an initial child of the awaited sign b of the immediately superordinate connected component a/b from a to b, it is hypothesized to do so and terminate recognition of the subordinate connected component, with x_t as the last observation of the terminated connected component (via production L+); or if the

[9] A deductive system consists of inferences or productions of the form: $\frac{P}{Q}R$, meaning premise P entails conclusion Q according to rule R.

observation can serve as a lower descendant of this awaited sign b, it is hypothesized to remain disjoint and form its own connected component (via production L–):

$$\frac{a/b \quad a''}{a/b''}\ b \rightarrow a''\ b'' \tag{L+}$$

$$\frac{a/b \quad a''}{a/b \quad a'/b''}\ b \xrightarrow{+} a'\ \ldots\ ;\quad a' \rightarrow a''\ b'' \tag{L–}$$

These initiation (F) and termination (L) productions are similar to the push and pop operations respectively of a nondeterministic pushdown automaton or the shift and reduce operations of a shift-reduce parser (taking all non-observation consequents in this model to be store items), except that the model observes a memory-based processing constraint that it may use no more than one initiation (F) and one termination (L) production per time step. When applied to parsing, this model is therefore more constrained than the PDA-based models of Crocker and Brants (2000) and Henderson (2004), which allow unlimited numbers of initiations and terminations between each pair of observations. An example derivation of the sentence The studio bought the publisher's rights, using these productions is shown in Figure 3.4.

Note that the derivation shown in Figure 3.4 contains only two instances of a connected component more than one word long being composed with a disjoint superordinate connected component, requiring temporal cueing or recall of the superordinate connected component (for example, this occurs after the genitive marker 's, merging the publisher's with the studio bought, and after the word rights, merging the sentence with the discourse context $\top/\top$). In general this only results from a non-initiation production (F–) followed by a termination production (L+). All other combinations of productions (non-termination in F–L– or F+L–, or initiation with immediate termination in F+L+) can be implemented by content-based sequential

cueing on the content of the preceding most subordinate state. This gives the system the capability to directly measure temporal cueing or recall operations during parsing. When extended to a probabilistic model, this measure can be expressed as a proportion of a distribution of weighted hypotheses, and potentially combined with other probabilistic measures like surprisal and entropy reduction (Hale, 2001, 2006).

Also note that if there is no bound on the number of disjoint connected-component states that can be hypothesized in a hierarchy, this system will be able to generate all and only those syntactic structures allowed by the context-free rules it is defined over. This can be shown by observing that (i) every binary syntactic tree over t observations must also contain t branches from a parent sign to a pair of children,[10] and (ii) for each observation x_t in a complete tree there is a unique largest sign which is co-final with x_t. Inspection of the F and L operations shows that F– and F+ isolate this largest co-final sign (as a in the consequent of F– or a' in the consequent of F+), and L+ and L– connect this sign (as a'') to a parent and a sibling in some tree (to b and b'' in L+ and to a' and b'' in L–). Since each such connection is unique, and since there are only t such connections in any complete tree, the system must be able to predict any complete tree for any sequence of t observations.

A probabilistic hierarchic sequence model

This process can be extended to calculate probabilities for syntactic tree structures by introducing probability distributions ϕ and λ over the binary F and L decisions defined above, with constraints taken directly from these productions. F decisions

[10]This includes a connection from the top-level sign of this structure to some discourse structure $\top$.

(about whether to initiate a new subordinate sequence) are constrained such that:

$$\mathsf{P}_{\phi}(\text{`}-\text{'} \mid b\ x) \neq 0 \quad \text{only if} \quad b \to x \tag{3.1a}$$

$$\mathsf{P}_{\phi}(\text{`}+\text{'} \mid b\ x) \neq 0 \quad \text{only if} \quad b \xrightarrow{+} a' ... \quad \text{and} \quad a' \to x \quad \text{for some } a' \tag{3.1b}$$

L decisions (about whether to terminate a subordinate sequence) are constrained such that:

$$\mathsf{P}_{\lambda}(\text{`}+\text{'} \mid b\ a'') \neq 0 \quad \text{only if} \quad b \to a''\ b'' \quad \text{for some } b'' \tag{3.2a}$$

$$\mathsf{P}_{\lambda}(\text{`}-\text{'} \mid b\ a'') \neq 0 \quad \text{only if} \quad b \xrightarrow{+} a' ... \quad \text{and} \quad a' \to a''\ b'' \quad \text{for some } a', b'' \tag{3.2b}$$

Constraints for probability distributions α and β over the active and awaited signs a and b in hypothesized connected-component states are also derived from F and L productions:

$$\mathsf{P}_{\alpha}(a' \mid b\ a'') \neq 0 \quad \text{only if} \quad b \xrightarrow{+} a' ... \quad \text{and} \quad a' \to a''\ b'' \quad \text{for some } b'' \tag{3.3}$$

$$\mathsf{P}_{\beta}(b'' \mid a'\ a'') \neq 0 \quad \text{only if} \quad a' \to a''\ b'' \tag{3.4}$$

These constraints are more precisely defined in the next section.

Since F productions take observations x as input and produce complete connected components a as output, and L productions take complete connected components a as input and produce incomplete connected-component states a/b as output, the process may only iterate by applying exactly one F production and one L production at each time step. Since F and L productions each have two (‘+’ and ‘–’) options, a complete hierarchic transition model σ can be defined using only four cases: one for each combination of F and L productions. These cases are represented as addends in the definition below. Each addend is a product of factors for: (i) hypothesizing a combination of an initiation and a termination of a subordinate state sequence

(using ϕ and λ probabilities), (ii) hypothesizing an active and awaited sign for the most subordinate connected-component state in the resulting hierarchy (using α and β probabilities), and (iii) deterministically carrying forward empty or unmodified states from the previous time step. All models depend on the depth d of the most subordinate connected-component state at the previous time step, and (in the case of the β model) on whether the first parameter is an active or awaited sign ('A' or 'B', respectively). In this definition, D is an arbitrary bound on the size of the state hierarchy (set to 4 in the evaluation described below), and $[\![...]\!]$ is a deterministic indicator function, evaluating to 1 if '...' is true, and 0 otherwise, used to represent a deterministic distribution.

The transition model is factored into three stages: σ, σ', and σ'', below. First, the probability is split across the two possible outcomes for the F decision — whether to introduce a new subordinate sequence or not — based on the most subordinate connected component state q^d_{t-1} in the state hierarchy:

$$\begin{aligned}\mathsf{P}_\sigma(q_t^{1..D} \mid q_{t-1}^{1..D}\ x_{t-1}) \stackrel{\text{def}}{=}\ & \mathsf{P}_{\phi_d}(\text{`}-\text{'} \mid b^d_{t-1}\ x_{t-1}) \cdot \mathsf{P}_{\sigma'_d}(q_t^{1..D} \mid q_{t-1}^{1..D}\ a^d_{t-1}) \\ & + \mathsf{P}_{\phi_d}(\text{`}+\text{'} \mid b^d_{t-1}\ x_{t-1}) \cdot \mathsf{P}_{\sigma'_{d+1}}(q_t^{1..D} \mid q_{t-1}^{1..D}\ x_{t-1}); \quad d \stackrel{\text{def}}{=} \max\{d' \mid q^{d'}_{t-1} \neq \text{`}-\text{'}\}\end{aligned} \tag{3.5a}$$

Then, for each F outcome, the transition model splits the remaining probability across the two possible outcomes for the L decision — whether to terminate the current most subordinate sequence or not — traversing the predicted tree downward from b^{d-1}_{t-1} if so (using rule $b^{d-1}_{t-1} \rightarrow a''\ b^{d-1}_t$), and traversing the predicted tree upward from a'' if not

(using rule $a_t^d \rightarrow a''\ b_t^d$):[11]

$$\mathsf{P}_{\sigma'_d}(q_t^{1..D} \mid q_{t-1}^{1..D}\ a'') \stackrel{\text{def}}{=} \mathsf{P}_{\lambda_d}(\text{`+'} \mid b_{t-1}^{d-1}\ a'') \cdot [\![a_t^{d-1} = a_{t-1}^{d-1}]\!] \cdot \mathsf{P}_{\beta_{\text{B},d-1}}(b_t^{d-1} \mid b_{t-1}^{d-1}\ a'') \cdot \mathsf{P}_{\sigma''_{d-1}}(q_t^{1..D} \mid q_{t-1}^{1..D})$$
$$+\ \mathsf{P}_{\lambda_d}(\text{`–'} \mid b_{t-1}^{d-1}\ a'') \cdot \mathsf{P}_{\alpha_d}(a_t^d \mid b_{t-1}^{d-1}\ a'') \cdot \mathsf{P}_{\beta_{\text{A},d}}(b_t^d \mid a_t^d\ a'') \cdot \mathsf{P}_{\sigma''_d}(q_t^{1..D} \mid q_{t-1}^{1..D}) \tag{3.5b}$$

Finally, for each combination of F and L outcomes, the transition model ensures the rest of the hierarchy $q_t^{1..D}$ is copied over from the previous time step, or replaced with null values below the most subordinate connected-component state in the hierarchy:

$$\mathsf{P}_{\sigma''_d}(q_t^{1..D} \mid q_{t-1}^{1..D}) \stackrel{\text{def}}{=} [\![q_t^{1..d-1} = q_{t-1}^{1..d-1}]\!] \cdot [\![q_t^{d+1..D} = \text{`–'}]\!] \tag{3.5c}$$

These transition probabilities σ are then combined with observation probabilities ξ to define a most likely sequence of connected component hierarchies $\hat{q}_{1..T}^{1..D}$:

$$\hat{q}_{1..T}^{1..D} \stackrel{\text{def}}{=} \operatorname*{argmax}_{q_{1..T}^{1..D}} \prod_{t=1}^{T} \mathsf{P}_{\sigma}(q_t^{1..D} \mid q_{t-1}^{1..D}\ x_{t-1}) \cdot \mathsf{P}_{\xi}(x_t \mid b_t^d); \quad d \stackrel{\text{def}}{=} \max\{d' \mid q_{t-1}^{d'} \neq \text{`–'}\} \tag{3.6}$$

This model predicts syntactic tree structure while restricting access to superordinate states as a memory-based recency constraint. Since the model is implemented as a simple sum of products, it is essentially equivalent to a localist representation in a recurrent neural network, albeit one with a hidden context unit for every combination of disjoint connected components, represented in an explicit state hierarchy. In this respect, the hierarchic sequence model described in this chapter is similar to the hierarchic connectionist memory model of Botvinick (2007). However, in order to maintain a close connection with linguistic notions of syntactic structure, the model is not trained using learning techniques traditionally applied to connectionist models.

[11] Note that the active sign of a superordinate state is deterministically carried forward whenever x_t is the last observation in a subordinate state sequence (when λ is positive). This is because the active sign of a superordinate state does not change when a subordinate state is terminated.

3.1.2 Application to probabilistic context free grammars

The constraints described in the previous section can be satisfied in a variety of ways. The model evaluated in this chapter is directly defined over a Probabilistic Context Free Grammar (PCFG), trained using latent variable induction (Petrov et al., 2006). PCFGs are widely used in parsing because they provide a simple branching stochastic process (Collins, 1997), because well-studied algorithms exist for inferring or refining PCFGs from data (Petrov et al., 2006), and because PCFGs have been shown to be useful as a basis for information-theoretic accounts of garden path effects and reading time delays (Jurafsky, 1996; Hale, 2001, 2003, 2006; Levy, 2008).

Side- and depth-specific rules

The general hierarchic sequence model described in the Model section can be defined for a given PCFG by first deriving side- and depth-specific rule probabilities and expected counts of initial descendant sub-signs derived from rule probabilities in Chomsky Normal Form (CNF):[12]

- $\mathsf{P}_{\gamma}(a \rightarrow x)$ denotes the probability of a sign of category a expanding into an observation,
- $\mathsf{P}_{\gamma_{s,d}}(a' \rightarrow a''\ b'')$ denotes the probability that a sign of category a' at hierarchy depth d occurring on initial (A) or final (B) side s of its parent will expand into an initial sign of category a'' followed by a final sign of category b'',
- $\mathsf{E}_{\gamma^*_d}(b \xrightarrow{+} a'\ ...)$ denotes the expected number of times a sign of category a' at hierarchy depth d will occur as an initial descendant sub-sign of another sign of category b, and

[12]PCFGs not in CNF can be compiled into CNF by binarizing with unique symbols.

- $\mathsf{E}_{\gamma^*_d}(b \xrightarrow{*} a' \ldots)$ denotes the expected number of times a sign of category a' at hierarchy depth d will occur either as equivalent to or as an initial descendant sub-sign of another sign of category b.

In lieu of a confusability model, the model instead imposes hard constraints on the number of distinct syntactically connected components allowed in each hypothesis. This has the effect of bounding the number of center embeddings allowed in any partial syntactic tree (in particular, initial children of final children in any CNF derivation). This is done by first computing a side- and depth-specific PCFG 'fit' model $\delta^{(i)}_{s,d}$, defining the probability that a subtree below a sign of category a, occurring on its parent's initial or final side $s \in \{A, B\}$, will fit within a bounded depth d of disjoint connected components. This fit model is computed according to the following recursive definition, where i is the recursive iteration:

$$\mathsf{P}_{\delta^{(0)}_{s,d}}(1 \mid a) \stackrel{\text{def}}{=} 0 \tag{3.7a}$$

$$\mathsf{P}_{\delta^{(i)}_{A,d}}(1 \mid a) \stackrel{\text{def}}{=} \sum_x \mathsf{P}_\gamma(a \to x) + \sum_{a',b'} \mathsf{P}_\gamma(a \to a'\ b') \cdot \mathsf{P}_{\delta^{(i-1)}_{A,d}}(1 \mid a') \cdot \mathsf{P}_{\delta^{(i-1)}_{B,d}}(1 \mid b') \tag{3.7b}$$

$$\mathsf{P}_{\delta^{(i)}_{B,d}}(1 \mid a) \stackrel{\text{def}}{=} \sum_x \mathsf{P}_\gamma(a \to x) + \sum_{a',b'} \mathsf{P}_\gamma(a \to a'\ b') \cdot \mathsf{P}_{\delta^{(i-1)}_{A,d+1}}(1 \mid a') \cdot \mathsf{P}_{\delta^{(i-1)}_{B,d}}(1 \mid b') \tag{3.7c}$$

Note that the only difference between the initial-sign ($\delta^{(i)}_{A,d}$) and final-sign ($\delta^{(i)}_{B,d}$) cases above is simply that the depth is incremented for initial children of final children. These initial-final zig-zags form the breaks between connected components in the hierarchy. In practice the recursive product is estimated to some constant I using value iteration (Bellman, 1957).

Now a side- and depth-specific PCFG model $\gamma_{s,d}$ can be defined by renormalizing over the probability mass isolated in $\delta^{(I)}_{s,d}$:

$$\mathsf{P}_{\gamma_{A,d}}(a \to a'\ b') \stackrel{\text{def}}{=} \frac{\mathsf{P}_{\gamma}(a \to a'\ b') \cdot \mathsf{P}_{\delta^{(I)}_{A,d}}(1 \mid a') \cdot \mathsf{P}_{\delta^{(I)}_{B,d}}(1 \mid b')}{\mathsf{P}_{\delta^{(I)}_{A,d}}(1 \mid a)} \tag{3.8a}$$

$$\mathsf{P}_{\gamma_{B,d}}(a \to a'\ b') \stackrel{\text{def}}{=} \frac{\mathsf{P}_{\gamma}(a \to a'\ b') \cdot \mathsf{P}_{\delta^{(I)}_{A,d+1}}(1 \mid a') \cdot \mathsf{P}_{\delta^{(I)}_{B,d}}(1 \mid b')}{\mathsf{P}_{\delta^{(I)}_{B,d}}(1 \mid a)} \tag{3.8b}$$

This renormalizing over $\delta^{(I)}_{s,d}$ ensures no probability mass is lost when the depth of the model is bounded. Again, the only difference between the initial- and final-sign cases is that the depth is incremented for initial children of final children.

Initial descendant sub-sign expected counts

The model also needs initial descendant sub-sign expected counts, which are based on the expected number of times a constituent of category a'' occurs at the beginning of a constituent of category b after any number of expansions. This is also estimated recursively with j as the recursive iteration:

$$\mathsf{E}_{\gamma^*_d}(b \xrightarrow{1} a'\ ...) \stackrel{\text{def}}{=} \sum_{b'} \mathsf{P}_{\gamma_{B,d}}(b \to a'\ b') \tag{3.9a}$$

$$\mathsf{E}_{\gamma^*_d}(b \xrightarrow{j} a''\ ...) \stackrel{\text{def}}{=} \sum_{a',b''} \mathsf{E}_{\gamma^*_d}(b \xrightarrow{j-1} a'\ ...) \cdot \mathsf{P}_{\gamma_{A,d}}(a' \to a''\ b'') \tag{3.9b}$$

$$\mathsf{E}_{\gamma^*_d}(b \xrightarrow{+} a''\ ...) \stackrel{\text{def}}{=} \sum_{j=1}^{\infty} \mathsf{E}_{\gamma^*_d}(b \xrightarrow{j} a''\ ...) \tag{3.9c}$$

$$\mathsf{E}_{\gamma^*_d}(b \xrightarrow{*} a''\ ...) \stackrel{\text{def}}{=} [\![b{=}a'']\!] + \mathsf{E}_{\gamma^*_d}(b \xrightarrow{+} a''\ ...) \tag{3.9d}$$

Here again, the recursive products and infinite sum are estimated to some constant J using value iteration (Bellman, 1957).

Transition operations for language comprehension

The model probabilities are then just straightforward probabilistic implementations of the constraints specified in Equations 3.1a–3.4 expressed in terms of side- and depth-specific rule probabilities and expected counts of initial descendant sub-signs from a PCFG γ:

1. The initiation model ϕ probabilities are calculated from the expected counts of a sign of syntactic category a' occurring as an initial descendant sub-sign of a larger sign of category b multiplied by the probability of that category generating an observation x:

$$\mathsf{P}_{\phi_d}(\text{`}-\text{'} \mid b\ a') \stackrel{\text{def}}{=} \frac{[\![b = a']\!] \cdot \sum_x \mathsf{P}_\gamma(a' \to x)}{\mathsf{E}_{\gamma_d^*}(b \stackrel{*}{\to} a'\ ...) \cdot \sum_x \mathsf{P}_\gamma(a' \to x)} \tag{3.10a}$$

$$\mathsf{P}_{\phi_d}(\text{`}+\text{'} \mid b\ a') \stackrel{\text{def}}{=} \frac{\mathsf{E}_{\gamma_d^*}(b \stackrel{+}{\to} a'\ ...) \cdot \sum_x \mathsf{P}_\gamma(a' \to x)}{\mathsf{E}_{\gamma_d^*}(b \stackrel{*}{\to} a'\ ...) \cdot \sum_x \mathsf{P}_\gamma(a' \to x)} \tag{3.10b}$$

2. The termination model λ probabilities are calculated from the expected counts of a sign of syntactic category a' occurring as an initial descendant sub-sign of a larger sign of category b multiplied by the probability of that sign having an initial child of category a'':

$$\mathsf{P}_{\lambda_d}(\text{`}+\text{'} \mid b\, a'') \stackrel{\text{def}}{=} \frac{\sum_{a',b''} [\![b = a']\!] \cdot \mathsf{P}_{\gamma_{B,d}}(a' \to a''\ b'')}{\sum_{a',b''} \mathsf{E}_{\gamma_d^*}(b \stackrel{*}{\to} a'\ ...) \cdot \mathsf{P}_{\gamma_{A,d}}(a' \to a''\ b'')} \tag{3.11a}$$

$$\mathsf{P}_{\lambda_d}(\text{`}-\text{'} \mid b\, a'') \stackrel{\text{def}}{=} \frac{\sum_{a',b''} \mathsf{E}_{\gamma_d^*}(b \stackrel{+}{\to} a'\ ...) \cdot \mathsf{P}_{\gamma_{A,d}}(a' \to a''\ b'')}{\sum_{a',b''} \mathsf{E}_{\gamma_d^*}(b \stackrel{*}{\to} a'\ ...) \cdot \mathsf{P}_{\gamma_{A,d}}(a' \to a''\ b'')} \tag{3.11b}$$

3. The active sign model α probabilities are calculated from the expected counts of a sign of syntactic category a' occurring as an initial descendant sub-sign of a larger sign of category b multiplied by the probability of that sign having an

initial child of category a'':

$$\mathsf{P}_{\alpha_d}(a' \mid b\, a'') \stackrel{\text{def}}{=} \frac{\sum_{b''} \mathsf{E}_{\gamma_d^*}(b \xrightarrow{+} a' \ ...) \cdot \mathsf{P}_{\gamma_{A,d}}(a' \rightarrow a''\ b'')}{\sum_{a',b''} \mathsf{E}_{\gamma_d^*}(b \xrightarrow{+} a' \ ...) \cdot \mathsf{P}_{\gamma_{A,d}}(a' \rightarrow a''\ b'')} \tag{3.12}$$

4. The awaited sign model β probabilities are simply the probability that a sign of category a has a final child of category b' given that it has an initial child of category a':

$$\mathsf{P}_{\beta_{s,d}}(b' \mid a\, a') \stackrel{\text{def}}{=} \frac{\mathsf{P}_{\gamma_{s,d}}(a \rightarrow a'\ b')}{\sum_{b'} \mathsf{P}_{\gamma_{s,d}}(a \rightarrow a'\ b')} \tag{3.13}$$

Transition model for language comprehension

These individual model probabilities are combined into a single hierarchic transition probability σ, as described in Equations 3.5a–3.5c of the previous section. These transition probabilities σ are then combined with preterminal and terminal probabilities π and ξ, described below.

In the previous section, observations were generated directly from awaited signs. In practice, it is more efficient to make the assumption that certain syntactic categories are preterminals, which generate a single lexical observation as a child. Such an assumption is made primarily for efficiency since only a subset of syntactic categories must then be considered for prediction, which reduces the complexity of the ϕ, α, and β models that depend on preterminal signs.

Formally, the preterminal probabilities π define the normalized probabilities of generating a preterminal p as an initial descendant sub-sign of a larger sign of category b:

$$\mathsf{P}_{\pi_d}(p \mid b) \stackrel{\text{def}}{=} \mathsf{E}_{\gamma_d^*}(b \xrightarrow{*} p \ ...) \cdot \sum_x \mathsf{P}_{\gamma}(p \rightarrow x) \tag{3.14}$$

Terminal probabilities ξ are then defined as the normalized probabilities of generating an observation x from a preterminal p:

$$\mathsf{P}_\xi(x \,|\, p) \stackrel{\text{def}}{=} \frac{\mathsf{P}_\gamma(p \rightarrow x)}{\sum_x \mathsf{P}_\gamma(p \rightarrow x)} \tag{3.15}$$

These transition probabilities σ, preterminal probabilities π, and terminal probabilities ξ are then combined to define a most likely sequence $\hat{q}_{1..T}^{1..D}$:

$$\hat{q}_{1..T}^{1..D} \stackrel{\text{def}}{=} \underset{q_{1..T}^{1..D}}{\text{argmax}} \prod_{t=1}^{T} \mathsf{P}_\sigma(q_t^{1..D} \,|\, q_{t-1}^{1..D}\, p_{t-1}) \cdot \mathsf{P}_{\pi_{d'}}(p_t \,|\, b_t^d) \cdot \mathsf{P}_\xi(x_t \,|\, p_t); \;\; d \stackrel{\text{def}}{=} \max\{d' \,|\, q_{t-1}^{d'} \neq \text{`\_'}\} \tag{3.16}$$

3.1.3 Evaluation

In order to determine whether the flattened hierarchy generated by the single initiation and single termination constraints described above could predict syntactic tree structures as accurately as a model without such constraints, the hierarchic sequence model described in this chapter was evaluated on a standard parsing task (Collins, 1997). This task is to reproduce labeled bracketings on a standard test set of newspaper text from the Wall Street Journal section of the Penn Treebank (Marcus et al., 1993), which can be directly compared against published results of other models. Comparable results to these systems would indicate that the memory-based constraints of the hierarchic sequence model do not harm language processing performance.

Training

The model evaluated in this chapter uses the Petrov et al. (2006) split-merge-smooth algorithm to extract a latent variable PCFG from the Penn Treebank (Marcus et al., 1993). The corpus delimits syntactic constituents with parentheses and

category labels. The split-merge-smooth algorithm attempts to find latent subcategorizations (splits) of each category label which conform to distinct distributions in a set of training data relative to the surrounding category labels. For example, the class of present tense ditransitive verbs (such as gives) may be discovered to have a different distribution than the class of present tense transitive verbs (such as owns), though both are typically labelled 'VBZ' (present tense verb) in the Treebank. Another iteration of the splitting algorithm may then find that certain categories appear more often as first arguments of ditransitive verbs (now that they have been distinguished) than as arguments of transitive verbs. Assigning each such distribution a unique subcategory label helps encode mild contextual information into each label. To avoid overfitting to the training data, splits which are not sufficiently statistically informative are then merged back into a larger category. The PCFG relations used in the previous section are then calculated over these refined grammar categories.

Results

The accuracy of the hierarchic sequence model as a parser was compared to that of the Petrov and Klein (2007) and Roark (2001) parsers using varying beam widths (numbers of competing hypotheses).[13] The Petrov and Klein (2007) parser is a state-of-the-art chart parser based on the same latent variable PCFG (Petrov et al., 2006) used to define the hierarchic sequence model evaluated in this chapter. As a chart parser, it does not calculate prefix probabilities like the model described in this chapter and therefore cannot be used to calculate complexity measures like surprisal or entropy reduction. The Roark (2001) parser is an incremental parser widely used

[13]The Petrov and Klein (2007) parser was run using the Viterbi decoding option on the latent variable grammar to be comparable with the other incremental parsers.

System	Precision	Recall	F-score
Roark 2001 (CNF)	86.6	86.5	86.5
Current Model (CNF, beam width 500)	86.6	87.3	87.0
Current Model (CNF, beam width 2000)	87.8	87.8	87.8
Current Model (CNF, beam width 5000)	87.8	87.8	87.8
Petrov Klein (CNF)	88.1	87.8	88.0
Petrov Klein (not CNF)	88.3	88.6	88.5
Dyer et al., 2016 (not CNF)			92.4

Table 3.1: Accuracy comparison with state-of-the-art syntactic parsers. Numbers in parentheses are the number of parallel activated hypotheses. The left-corner parser used here restricts trees to Chomsky Normal Form (CNF), in which trees are binary branching at all nonterminals except preterminals. This makes the model less able to reproduce unary branches in the Penn Treebank.

in cognitive modeling evaluations, and can be used to calculate prefix probabilities necessary for calculating complexity measures like surprisal, but it is not as accurate as that of Petrov and Klein. Neither the Petrov and Klein nor Roark parsers are defined in terms analogous to sequentially or temporally cued recall operations.

The evaluated hierarchic sequence model restricts the number of disjoint connected components in any hypothesis to at most four. This limit was empirically determined to be sufficient to achieve greater than 99.9% coverage on the Wall Street Journal Corpus (Schuler et al., 2010).

All parsers were trained on Sections 02-21 of the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993) and tested on Section 23, following the standard parsing evaluation. No tuning was done as part of the conversion to a sequence model. With the exception of the Roark (2001) parser,[14] all parsers used 5 iterations of the

[14]The top-down nature of the Roark (2001) parser is not amenable to efficient use of the subcategorizations output by the split-merge-smooth algorithm.

Petrov et al. (2006) split-merge-smooth algorithm.[15] The results are shown in Table 3.1. Note that the Petrov and Klein (2007) parser allows unary branching within the syntactic tree structure, which is not directly supported by the set of production rules described in the model section of this chapter. To obtain a fair comparison, it was also run with strict binarization (restricting the grammar to Chomsky Normal Form). The hierarchic sequence model described in this chapter achieves comparable accuracy to the Petrov and Klein (2007) parser, assuming a strictly binary-branching syntax tree, and superior accuracy to the Roark (2001) parser.

3.2 Complexity measures

Having demonstrated the accuracy of the parser in the preceding section, this section outlines a number of incremental psycholinguistic complexity measures which are output by the parser.

3.2.1 PCFG surprisal

Hale (2001) demonstrated that information-theoretic surprisal (Shannon, 1948; Attneave, 1959) could predict the difficulty difference between subject and object relative clauses and the difficulty associated with garden path sentences. That is, the conditional log probability of a word given the preceding context is proportional to the difficulty people experience during incremental sentence processing. Levy (2008) showed that this relationship is equivalent to a model that predicts difficulty according to the amount of probability mass that needs to be reallocated after each new lexical observation. That is, a probabilistic fully parallel incremental parser needs to estimate

[15]This is the recommended number of split-merge iterations to obtain high accuracy while avoiding overfitting (Petrov and Klein, 2007).

the probability of each parse hypothesis after each new observation. The amount of probability mass that is reallocated during this processing step is equivalent to the conditional log probability of the word given its context (i.e. the word's surprisal, S). Roark et al. (2009) demonstrated that computation of surprisal over an incremental beam (B_t) provides a close approximation to the true surprisal of a word (w_t) given the preceding sentential context ($w_{1..t-1}$) because the extremely low probability events that are excluded from a beam approximation will not contribute much probability mass to the surprisal measure.

$$S(w_t, w_{1..t-1}) = -\log \left[\sum_{q_t^{1..D}, p_t \in B_t} \mathrm{P}(w_t \mid p_t) \cdot \mathrm{P}(q_t^{1..D}, p_t \mid w_{1..t-1}) \right] \tag{3.17}$$

Surprisal and corrections for using surprisal to predict reading times are explored further in Chapter 4.

3.2.2 Incremental memory load

It has long been observed that center-embedded sentences cause difficulty during sentence processing (Yngve, 1960; Miller and Chomsky, 1963). Psycholinguistic theories have attempted to capture this effect by predicting longer reading times for embeddings that intervene in an incomplete dependency (Gibson, 2000) or for embeddings that contain material with confusable syntactico-semantic features (Lewis and Vasishth, 2005). Some studies have suggested that the difficulty of processing embedded structures may simply be driven by the infrequency of deeply embedded linguistic phrases (Hale, 2001; Frank et al., 2016), though Phillips (2010) has hypothesized that grammar rule probabilities may be grounded in memory limitations.[16]

[16]Frank et al. (2016) attempt to guard against this possibility by showing that individual subjects are more or less susceptible to supposed memory limitations depending on the language being processed. However, their findings don't completely rule out the possibility that people may disprefer

Regardless of a particular theory about where the processing difficulty comes from, embedded structures are of interest due to the difficulty they are associated with.

A center embedding is defined as any left syntactic branch from any right syntactic branch. The center embedding is concluded when the rightmost child of the subtree introduced by that left branch has been processed. The parser described in this chapter tracks the embedding depth of each parse hypothesis ($q_t^{1..D}$) separately for each new word (w_i) and uses these to compute a weighted embedding depth (D) where the depth of each hypothesis is weighted by its probability:

$$D(w_t) = \sum_{q_t^{1..D}, p_t \in B_t} \mathrm{P}(q_t^{1..D}, p_t, w_t \mid w_{1..t-1}) \cdot d; d \stackrel{\text{def}}{=} \max\{d' \mid q_t^{d'} \neq \text{`–'}\} \tag{3.18}$$

Embedding depth will not be explored further in this thesis.

3.2.3 Embedding Difference

Wu et al. (2010) hypothesized that the process of increasing the working memory load might be costly, and formalized this cost as embedding difference:

$$\Delta_D(w_t) = D(w_t) - D(w_{t-1}) \tag{3.19}$$

They showed that such a measure corresponds to a kind of depth-weighted surprisal, so essentially this measure combines a cost for memory load with a cost for infrequency of an observation. They demonstrated that such a measure could effectively predict self-paced reading times, and a similar finding was obtained by van Schijndel and Schuler (2013) on the Dundee eye-tracking corpus (Kennedy et al., 2003). Embedding difference will not be explored further in this thesis.

embedded constructions due to memory limitations and that this dispreference further exaggerates the processing complexity of embedded constructions by making them less frequent. Such an effect could be differentially distributed across languages, explaining the cross-linguistic difference observed by Frank et al. (2016).

Note that this measure predicts that readers will slow down when increasing memory load and will speed up when decreasing memory load. In other words, this measure predicts faster reading times at embedding-reduction sites, a somewhat controversial hypothesis (see, e.g., Gibson, 2000). Recently, Shain et al. (2016) used a very large (181 subjects) self-paced reading corpus (Natural Stories; Futrell et al., in prep), and found that there is a small but robust slowing of reading times at integration sites, suggesting previous results showing faster reading at embedding-reduction sites may have been false positives.

3.2.4 Entropy reduction

Information-theoretic entropy (H; Shannon, 1948) represents the uncertainty of a random process (X) based on the possible values of that process:

$$H(x) = -\sum_{x \in X} \mathrm{P}(x) \log \mathrm{P}(x) \tag{3.20}$$

Wilson and Carroll (1954) defined the informational content of words via the concept of entropy reduction, which corresponds to the degree to which an observation (w_t) reduces uncertainty about upcoming observations compared to the preceding observation (w_{t-1}):

$$\Delta_H(w_t) = \max(0, H_{t-1} - H_t) \tag{3.21}$$

Hale (2006) extended this concept with probabilistic context-free grammars and coined the entropy reduction hypothesis, which claims that 'entropy reduction is positively related to human sentence processing difficulty.' Intuitively, this means that people will have greater processing difficulty when an observation has a greater informational load.

Hale (2006) computes this measure over all possible derivations of a word sequence, given a small grammar. However, this measure would ideally be extended to large corpora with large grammars, which makes it impractical to compute entropy over all possible derivations. Instead, this work follows Wu et al. (2010) who demonstrated that entropy reduction also correlates with self-paced reading times when the entropy refers to uncertainty about the correct parse of previous observations. That is, the parser measures the entropy of the hypothesis beam at each time step (B_t) over all component hypotheses ($q_t^{1..D}$) and reports how much each new observation (w_t) reduces the beam's entropy:

$$\Delta_H(w_t) = \max(0, H_{t-1} - H_t) \tag{3.22}$$

$$= \max(0, H(w_{t-1} \mid w_{1..t-2}) - H(w_t \mid w_{1..t-1})) \tag{3.23}$$

$$= \max(0, \sum_{q_{t-1}^{1..D} \in B_{t-1}} \mathrm{P}(q^{1..D}, w_{t-1} \mid w_{1..t-2}) \log \mathrm{P}(q^{1..D}, w_{t-1} \mid w_{1..t-2}) - \sum_{q_t^{1..D} \in B_t} \mathrm{P}(q^{1..D}, w_t \mid w_{1..t-1}) \log \mathrm{P}(q^{1..D}, w_t \mid w_{1..t-1})) \tag{3.24}$$

Entropy reduction will not be explored further in this thesis though entropy, itself, and approximations to estimate entropy are explored in Chapter 5.

3.3 Conclusion and discussion

This chapter has shown that the flattened hierarchic sequence model described herein can obtain similar accuracy to a well-regarded non-psycholinguistic parser (the Petrov and Klein, 2007, parser). Therefore, the seemingly austere constraints of shallow hierarchic sequential prediction do not harm performance. The ready application of such general prediction to syntactic parsing suggests that human language processing might be performed using a shallow hierarchic sequential process similar to

those described in existing computational models of memory (Howard and Kahana, 2002; Botvinick, 2007).

Some combinations of operations in this model correspond to combinators in a Combinatory Categorial Grammar (CCG) (Steedman, 2000) in a maximally incremental parse. In particular:

- F+ followed by L– performs the CCG operation of forward type raising of x_t,
- F+ followed by L+ performs forward type raising of x_t followed by the CCG operation of forward function composition of a/b on this raised category, and
- F– performs the CCG operation of forward function application of a/b on x_t.

The model described in this chapter can therefore be taken as an exploration of the origins of combinators as a consequence of sequential and temporal cueing operations.[17]

Sturt and Lombardo (2005) warn about the need for constituents to be interconnected in processing, in order to pass along appropriate information to predict reading time delays. But the definition of connectivity they use includes underspecified or non-immediate relations such as the initial descendant sub-sign relations described in this chapter, and is therefore more permissive than the immediate graph-theoretic notion of connectivity used to define the limits of connected components. In the

[17]However, nothing in this account predicts any of the more specialized combinators (like backward cross composition, involved in right-node raising and filler-gap constructions with non-peripheral gaps), or the constrained set of elementary categories (S and NP signifying truth values and entities) normally associated with CCG. Moreover, although recall of superordinate connected components is dispreferred due to recency effects, nothing in this account rules out conditional dependencies in productions that cross multiple connected components in the hierarchy, which would violate the CCG principle of adjacency for combinators, providing the original motivation for some of the more sophisticated CCG combinators such as backward cross composition. A more thorough exploration of some of these issues would be an interesting topic for future research.

model described in this chapter, all disjoint connected components within the same hypothesis are syntactically connected by Sturt and Lombardo's (2005) more permissive definition since the awaited signs of superordinate connected components are related by initial descendant sub-sign relations ($\xrightarrow{+}$) to the active signs of subordinate connected components. Indeed, probabilistic operations can be defined to be dependent on other connected components in this model as well (although recall of superordinate connected components is dispreferred due to recency effects).

PCFGs provide a simple branching probabilistic model that can be used to generate complex syntactic structures and can be trained to predict these structures for novel sentences with high accuracy. This chapter shows that this kind of model can be incrementally processed in a straightforward way that is compatible with current assumptions about working memory. Carefully trained but computationally simple models like this may provide a framework for evaluating the contribution of recency and other memory-based effects on processing complexity (for example, costs associated with F– or L+ operations) on top of frequency effects currently measured by probabilistic measures like surprisal, entropy reduction, and uniform information density.

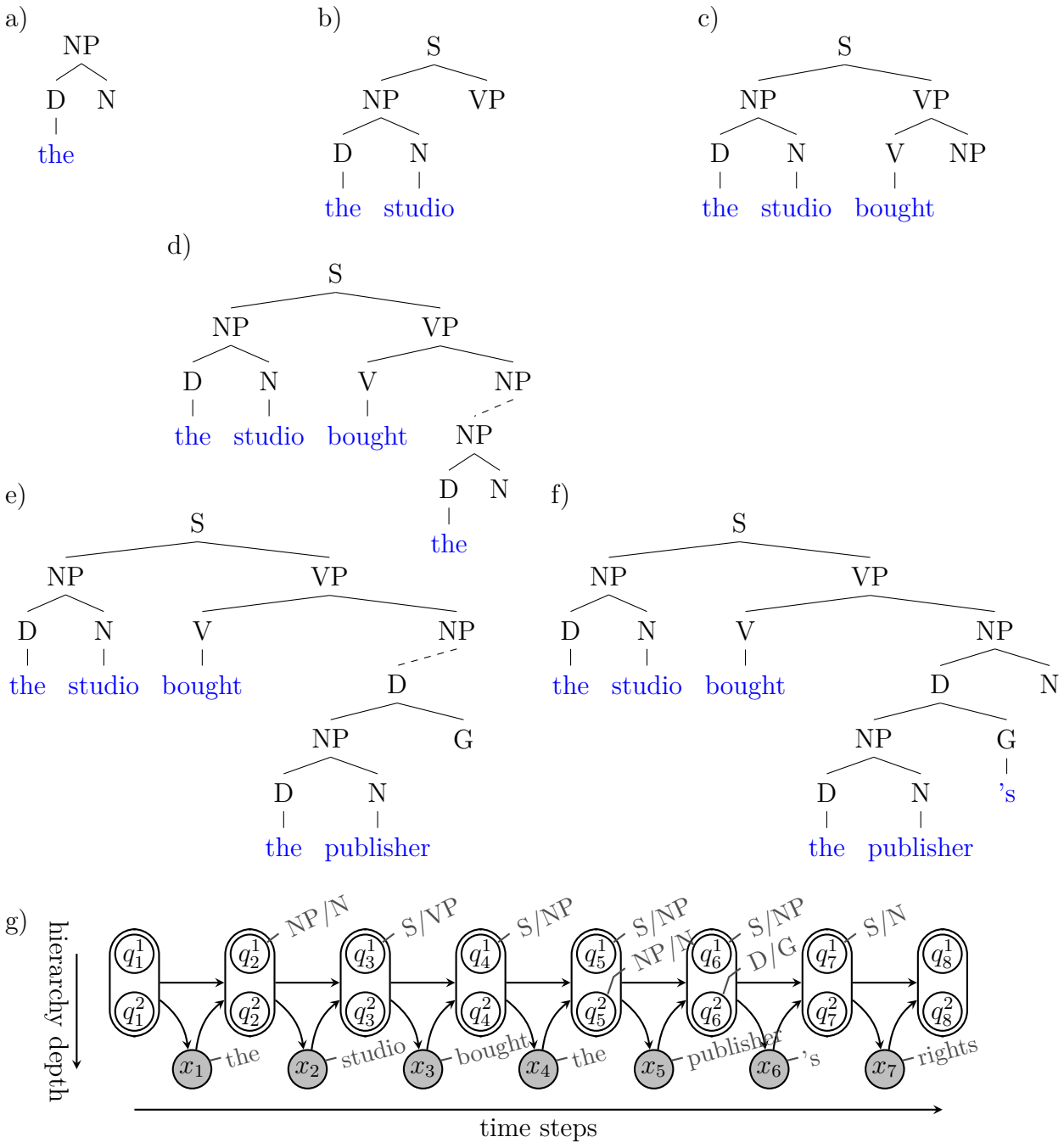

Figure 3.2: Incrementally constructed representations of the syntactic structure of the sentence The studio bought the publisher's rights (a–f), and the associated sequence of random variable values in a hierarchic sequential prediction model (g). Open circles represent hidden variables, shaded circles represent observed variables (x_t), and directed edges represent conditional dependencies. 'Pea-pod' ovals summarize dependencies over subsumed variables. Selected random variables are also annotated with example values, shown diagonally.

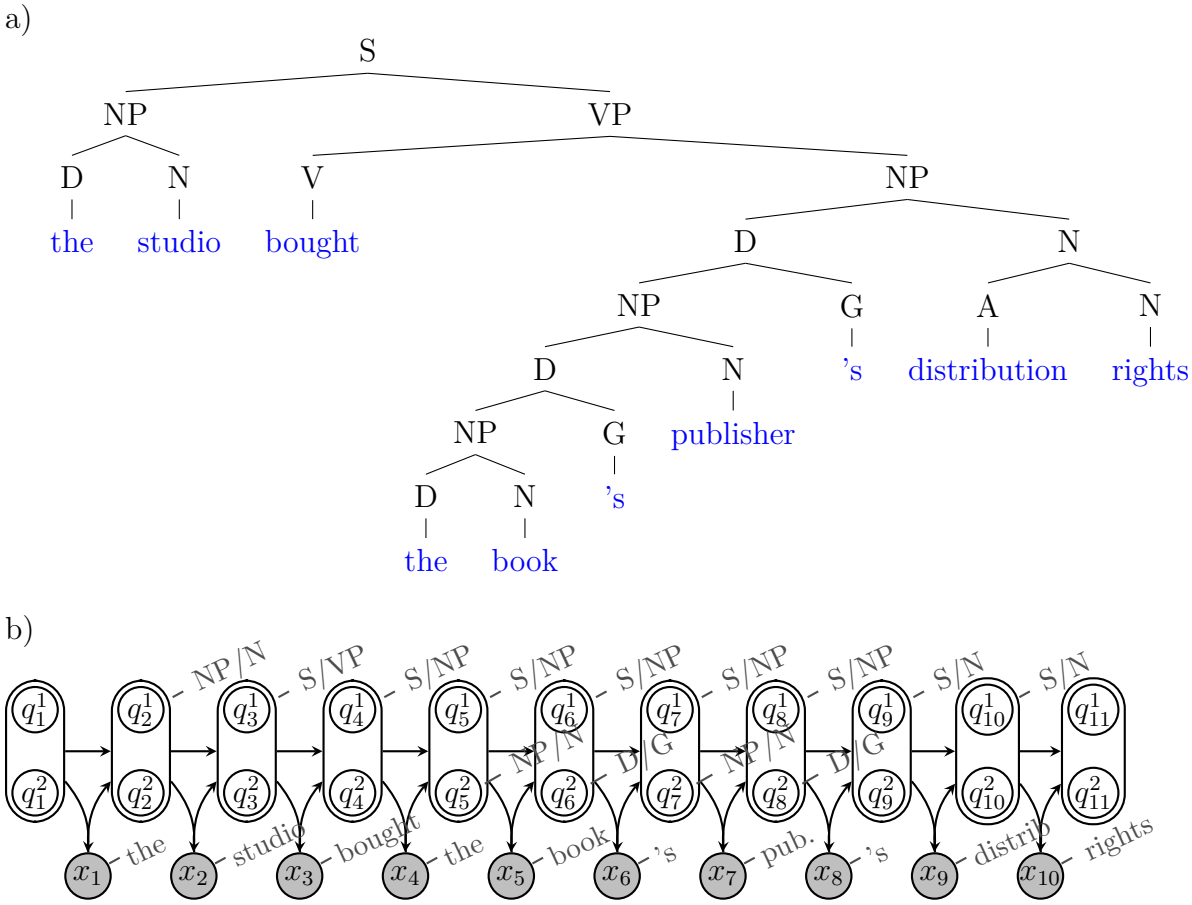

Figure 3.3: Phrase structure tree for the sentence The studio bought the book's publisher's distribution rights (a), with repeated initial and final signs for book's and distribution; and the associated sequence of random variable values in a hierarchic sequential prediction model (b), with corresponding repeated states D/G and S/N at time steps 8 and 10.

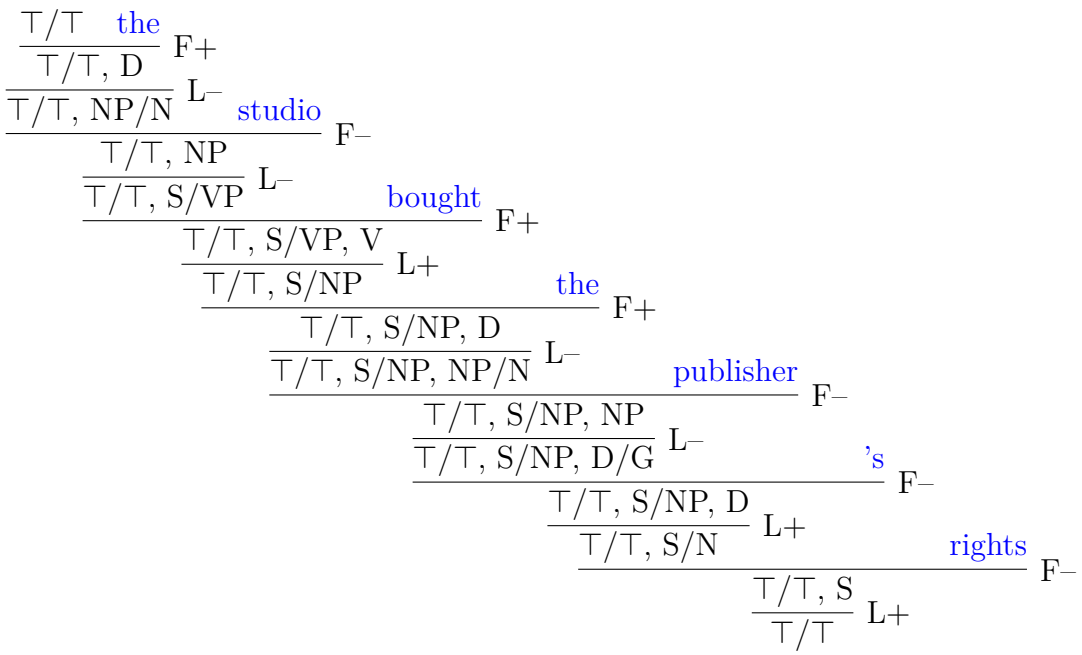

Figure 3.4: Derivation of the sentence The studio bought the publisher's rights, using F+, F–, L+, and L– productions.

Chapter 4: The influence of surprisal on reading times*

This chapter identifies and corrects an inconsistency in how conditional probabilities are used to predict reading times. Specifically, surprisal is conditioned on adjacent lexical context, but reading involves saccades between non-adjacent material. Therefore readers may respond behaviorally to complexity in the intervening material that is not modeled by traditional surprisal when it is used to predict complexity of spans larger than one item (e.g., first pass reading times). Summing surprisal over each saccade region improves the fit of n-gram surprisal to reading times, which makes probabilistic context-free grammar (PCFG) surprisal less effective as a reading time predictor, though it remains predictive in some corpora.

4.1 Introduction

Rare words and constructions produce longer reading times than their more frequent counterparts. Such effects can be captured by n-gram and probabilistic context-free grammar (PCFG) surprisal. Surprisal theory (Hale, 2001; Levy, 2008) predicts reading times will be directly proportional to the amount of new information which

*Parts of this chapter were originally published in van Schijndel and Schuler (2015) and in van Schijndel and Schuler (2016).

must be processed, as calculated by a generative model. However, the surprisal measures commonly used in eye-tracking studies (e.g., Demberg and Keller, 2008; Demberg et al., 2013; Frank, 2017) omit probability estimates for words skipped in saccades. Therefore, the generative model assumed by those studies does not account for the information contributed by the skipped words even though those words must be processed by readers. This chapter uses two reading time corpora to show that cumulative n-gram probabilities significantly improve an n-gram baseline to better capture the correlation between sequential frequency statistics and reading times. The ability of PCFG surprisal to predict reading times is shown to be much weaker given the improved n-gram measures, and this chapter suggests other measures may be necessary to adequately estimate the impact of syntactic processing on reading times.

First, this work uses a stronger n-gram baseline than that used in previous studies by replacing a bigram baseline computed from 101 million words with a back-off 5-gram baseline computed over 2.96 billion words. Second, while previous work has used the surprisal from the end of each eye-movement region to model reading times in that region, this chapter defines two cumulative surprisal measures (n-gram surprisal in Section 4.2; PCFG surprisal in Section 4.3), which are mathematically better able to estimate the amount of information conferred by each new region. Section 4.4 introduces the two corpora used in this chapter. Section 4.5 describes the linear mixed modeling baseline predictors and techniques used in this chapter. Section 4.6 shows that cumulative n-gram surprisal is better correlated with reading times than non-cumulative n-gram surprisal. Section 4.7 finds that cumulative PCFG surprisal is less effective than non-cumulative PCFG surprisal at predicting reading times, though

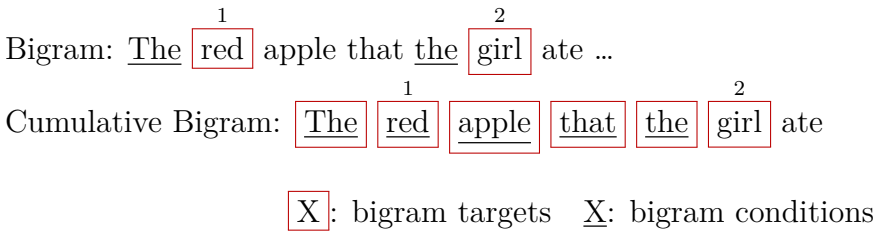

Figure 4.1: Eye movements jump between non-adjacent fixation regions (1, 2), while traditional n-gram measures are conditioned on the preceding adjacent context, which is never generated by the typical surprisal models used in eye-tracking studies. Cumulative n-grams sum the n-gram measures over the entire skipped region in order to better capture the information that readers need to process.

PCFG surprisal generally becomes much less effective as a reading time predictor with the n-gram improvements introduced in this chapter. Section 4.8 shows that at least some of the remaining predictivity of PCFG surprisal may be attributed to non-local syntactic dependencies.

4.2 Cumulative n-gram surprisal

N-gram surprisal is conditioned on the preceding context (see Equation 4.1). As stated in the introduction, however, direct use of this factor in a reading time model ignores the fact that some or all of the preceding context may not be generated if the associated lexical targets were not previously fixated by readers (see Figure 4.1). The lack of a generated condition results in a probability model that does not reflect the influence of words skipped during saccades. This deficiency can be corrected by accumulating n-gram surprisal over the entire saccade region (see Equation 4.2).

$$\text{n-gram}(w, i) = -\log \text{P}(w_i \mid \ w_{i-n} \ldots w_{i-1}) \tag{4.1}$$

Factors	Durations	
	R_4^4	R_5^6
Bigram Surprisal	$-\log \mathrm{P}(w_4 \mid w_3)$	$-\log \mathrm{P}(w_6 \mid w_5)$
Cumulative Bigram Surprisal	$-\log \mathrm{P}(w_4 \mid w_3)$	$\sum_{i=5}^{6} -\log \mathrm{P}(w_i \mid w_{i-1})$

Table 4.1: Bigram factors and their predictions of reading times in example eye-tracking regions. w_i represents word i. R_i^j represents the region from w_i to w_j (inclusive).

$$\text{cumu-n-gram}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} -\log \mathrm{P}(w_i \mid \ w_{i-n} \ldots w_{i-1}) \tag{4.2}$$

where w is a vector of input tokens, f_{t-1} is the index of the previous fixation, f_t is the index of the current fixation.

As a motivating example of the utility of cumulative n-grams, consider Table 4.1. The standard bigram factor (top line) predicts that the reading time of the region that ends with word 6 depends on word 5, but the probability of word 5 given its context is never included in the model, so an improbable transition between words 4 and 5 would not be caught. This might allow another factor to inappropriately receive credit for an extra long reading duration for the region ending with word 6. Instead, a better model would include the probabilities of every word in the sequence since that is the information that will need to be processed by the reader (see Figure 4.1 for an example).

4.3 Cumulative PCFG Surprisal

Probabilistic context-free grammar (PCFG) surprisal is similar to n-gram surprisal in that it is also conditioned on preceding context, but PCFG surprisal is conditioned on hierarchic structure rather than on linear lexical sequences (see Equation 4.3).

Factors	Durations	
	R_4^4	R_5^6
PCFG	$-\log \mathrm{P}(T_4 = w_4 \mid T_{1..3} = w_{1..3})$	$-\log \mathrm{P}(T_6 = w_6 \mid T_{1..5} = w_{1..5})$
Cumu-PCFG	$-\log \mathrm{P}(T_4 = w_4 \mid T_{1..3} = w_{1..3})$	$\sum_{i=5}^{6} -\log \mathrm{P}(T_i = w_i \mid T_{1..i-1} = w_{1..i-1})$

Table 4.2: PCFG surprisal factors and their predictions of reading times in example eye-tracking regions. w_i represents word i, T is a random variable over syntactic trees, and T_i is a terminal symbol in a tree. R_i^j represents the region from w_i to w_j (inclusive).

PCFG surprisal, therefore, suffers from the same deficiency as non-cumulative n-gram surprisal when modeling reading times: the condition context is never generated by the model. As with cumulative n-grams, cumulative PCFG surprisal of a region can be calculated by simply summing the PCFG surprisal of each word in the region (see Equation 4.4).

$$\mathrm{PCFG}(w, i) = -\log \mathrm{P}(T_i = w_i \mid \ T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1}) \tag{4.3}$$

$$\text{cumu-PCFG}(w, f_{t-1}, f_t) = \sum_{i=f_{t-1}+1}^{f_t} -\log \mathrm{P}(T_i = w_i \mid \ T_1 \ldots T_{i-1} = w_1 \ldots w_{i-1}) \tag{4.4}$$

where w is a vector of input tokens, f_{t-1} is the index of the previous fixation, f_t is the index of the current fixation, T is a random variable over syntactic trees and T_i is a terminal symbol in a tree. In psycholinguistic models, PCFG surprisal estimates the influence of incremental hierarchic context when processing a given word. For examples of these measures, see Table 4.2. A non-cumulative total PCFG surprisal factor (top line) would predict that duration of region R_5^6 depends on T_5 (the set of trees that can span from w_1 to w_5), but the probability of generating the prefix of T_5 is never fully calculated by this factor.

4.4 Data

This work makes use of the Dundee (Kennedy et al., 2003) and University College London (UCL) (Frank et al., 2013) eye-tracking corpora. The Dundee corpus consists of eye-tracking data from 10 subjects who read 2388 sentences of news text from the newspaper, The Independent. In contrast, the UCL corpus has reading time data from 43 subjects who read 361 sentences drawn from a series of self-published online novels. The articles in the Dundee corpus were read in their entirety with comprehension questions after each article. The sentences in the UCL corpus were presented isolated from one another, in a random order, with sentence-level comprehension questions appearing after half of the sentences.

The Dundee corpus and half of the sentences in the UCL corpus (every other sentence) were used for exploratory analyses, while the rest of the UCL corpus was set aside for significance testing.[18]

The corpora were parsed using the van Schijndel et al. (2013a) left-corner parser described in Chapter 3, which outputs a wide variety of incremental complexity measures computed during parsing. 5-gram back-off n-gram probabilities were computed for each word using the KenLM toolkit (Heafield et al., 2013) trained on Gigaword 4.0 (Graff and Cieri, 2003). Models were fit to Box-Cox transformed first-pass reading times for all experiments in this chapter ($\lambda \approx 0.02$; Box and Cox, 1964).[19] Fixation data was excluded from analysis if the fixation occurred on the first or last word of

[18]The different treatment of the two corpora arises from the fact that the Dundee analyses (except the future surprisal analyses) were originally reported first in van Schijndel and Schuler (2015), while the UCL analyses were conducted later with an eye to improving statistical practices and were reported in van Schijndel and Schuler (2016).

[19]The Box-Cox transform helps make the distribution of reading times more normal. The results here were also replicated using non-transformed reading times.

a sentence or line or if it followed an unusually long saccade, defined here and in previous work (Demberg and Keller, 2008) as a saccade over more than 4 words (3% of the Dundee corpus; 2.5% of the UCL corpus).

4.5 Modeling

All evaluations are done with linear mixed effects models[20] using lme4 (version 1.1-7; Bates et al., 2014).[21] There are two dependent reading time variables of interest in this study: first pass durations and go-past durations.[22] During reading, a person's eye can jump over multiple words each time it moves, this study refers to each new span of words as a region. First pass durations measure elapsed time until a person's eye leaves a given region. Go-past durations measure elapsed time until a person's eye moves further in the text. For example, in the fixation sequence: word 4, word 6, word 3, word 7, the first region would be from word 4 to word 6 and the second region would be from word 6 to word 7. The first pass duration for the first region would consist of the time fixated on word 6 before leaving the region for word 3, while the go-past duration would consist of the duration from the fixation of word 6 until the fixation of word 7. Separate models are fit to each centered dependent variable.

[20]Linear mixed modeling is a linear regression technique that separately estimates the variance for generalizable (fixed) population-level factors (e.g., human sensitivity to word length) and for non-generalizable (random) factors (e.g., each subject's individual sensitivity to word length).

[21]The models are fit using both the default bobyqa and the gradient nlminb algorithms to work around convergence issues. If a bobyqa model's final relative gradient during fitting is greater than 0.002, the model fit is compared across both fitting algorithms to determine whether convergence occurred. If the model has not converged, the model is refit without random intercept correlations. Subsequently, the random slopes of baseline factors are removed one-at-a-time until convergence is achieved.

[22]Both dependent measures were used for the Dundee analyses, but only first pass durations were used as dependent variables for the UCL analyses.

There are a number of independent variables in all evaluations in this study: sentence position (sentpos), word length (wlen), region length in words (rlen), whether the previous word was fixated (prevfix), basic 5-gram surprisal of the current word given the preceding context (5-gram), and cumulative 5-gram surprisal over the region (cumu-5-gram). All independent predictors are centered and scaled before being added to each model.

Each mixed effects model contains random intercepts for subject, word, and sentence ID crossed with subject. The latter helps account for the fact that multiple events occur within a single sentence/subject pairing, so those events are not independent. The models contain random by-subject slopes for all fixed effects. The following evaluations use ablative testing to determine whether a fixed effect significantly improves the fit of a model compared to a model without that fixed effect. All models in a given evaluation include random slopes for all fixed effects used in that evaluation, even if the fixed effect is absent from that particular model.

4.6 Experiment 1: Cumulative N-gram Surprisal in Reading Times

To test the effectiveness of using cumulative n-gram factors to predict reading times on the Dundee corpus (Table 4.3), the baseline omits the fixed n-gram factors. Then, the same model is fit to reading times after adding just a fixed effect for n-gram and after adding just a fixed effect for cumulative n-gram. Finally, a model

Model	First Pass		Go-Past	
	Log-Likelihood	AIC	Log-Likelihood	AIC
Baseline	−1212399	2424868	−1261582	2523234
Base+N-gram	−1212396†	2424864	−1261577*	2523226
Base+Cumu-N-gram	−1212392*	2424856	−1261576*	2523224
Base+Both	−1212387*	2424848	−1261570*	2523214

Baseline random slopes: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram
Baseline fixed effects: sentpos, wlen, rlen, prevfix

Table 4.3: Goodness of fit of n-gram models to reading times in the Dundee corpus.[25] Significance testing was done between each model and the models in the section above it. Significance for Base+Both applies to improvement over each of the n-gram models. † $p < .05$ * $p < .01$

is fit with both the cumulative and non-cumulative n-gram factors as fixed effects.[23] Significance between the models is determined using likelihood ratio testing.[24]

Table 4.3 shows that both n-gram factors significantly improve the fit of the model and the final line shows that each factor provides a significant orthogonal improvement. Both n-gram factors were therefore included as fixed effects and as by-subject random slopes in the baselines of the remaining Dundee analyses.

The same four models were fit to the UCL corpus (Table 4.4), and cumulative 5-grams again provided a significant improvement over basic n-grams ($p < 0.001$), but unlike with the Dundee analysis, basic n-grams do not improve over cumulative

[23]To ensure effects are not driven by individual subject differences, by-subject random slopes for both predictors of interest are included in all models. This practice is maintained throughout this thesis.

[24]Twice the log-likelihood difference of two nested models can be approximated by a χ^2 distribution with degrees of freedom equal to the difference in degrees of freedom of the models in question. The probability of obtaining a given log-likelihood difference D between the two models is therefore analogous to $P(2 \cdot D)$ under the corresponding χ^2 distribution.

[25]Log-likelihood values are rounded to the nearest whole number, which is why the difference between Base and Base+Both can be larger than the cumulative difference between Base and the other two models.

Model	N-gram vs Cumu-N-gram		
	$\hat{\beta}$	Log-Likelihood	AIC
Baseline		-12702	25476
Base+Basic	0.035	-12689^*	25451
Base+Cumulative	0.055	-12683^*	25440
Base+Both		-12683^*	25442

Baseline random slopes: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram
Baseline fixed effects: sentpos, wlen, rlen, prevfix

Table 4.4: Goodness of fit of n-gram models to first pass reading times in the UCL corpus. Significance testing was performed between each model and the models in the section above it. Significance for the Base+Both model applies to its improvement over the Base+Basic model. * $p < .001$

n-grams on this corpus ($p > 0.05$). The unreliability of non-cumulative n-grams as a reading time predictor compared to the cumulative variant suggests psycholinguists should use cumulative n-grams as a lexical frequency control in their reading time experiments. Further, the benefit of cumulative n-grams suggests that the lexical processing of words skipped during a saccade has a time cost similar to directly fixated words.[26]

4.7 Experiment 2: Cumulative PCFG Surprisal in Reading Times

The preceding section showed that applying region accumulation to an n-gram factor improves a model's fit to reading times. Previous work suggests region accumulation might improve the fit of syntactic factors to reading times (van Schijndel and Schuler, 2013; van Schijndel et al., 2013b), but the baselines in those studies only

[26]It is important to note that, in using first pass and go past durations as the dependent variables, it is possible that the words skipped by the initial saccade were fixated during a first-pass or go-past regression. This possibility seems unlikely to be so widespread and consistent as to drive these findings, but it is left to future work to test this possibility.

Model	First Pass		Go-Past	
	Log-Likelihood	AIC	Log-Likelihood	AIC
Baseline	−1212260	2424627	−1261488	2523084
Base+PCFG	−1212253*	2424617	−1261481*	2523072
Base+CumuPCFG	−1212259	2424627	−1261487	2523085
Base+Both	−1212253*	2424619	−1261481*	2523073

Baseline random slopes: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram, PCFG, cumuPCFG; Baseline fixed effects: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram.

Table 4.5: Goodness of fit of hierarchical syntax models to reading times in the Dundee corpus. Significance testing was done between each model and the models in the section above it. Significance for Base+Both applies only to improvement over the CumuPCFG model. * $p < .01$

included unigram and bigram statistics and did not apply region accumulation to the n-gram models. It does make intuitive sense that region accumulation would help improve the fit of total PCFG surprisal for the same reason accumulating n-grams helps.

As in the previous section, a baseline model is fit to Dundee reading times without a fixed effect for surprisal, then surprisal is added as a fixed effect and significance of the fixed effect is determined using a likelihood ratio test with the baseline. The results (Table 4.5) show that PCFG surprisal is a significant predictor of both first pass and go-past durations even over a strong baseline including both types of n-gram factors.

When tested, however, the present work does not find any improvement from region accumulation of PCFG surprisal when stronger n-gram factors are also included (Table 4.5, Row 2), suggesting that the improvement in previous studies may have been due to latent n-gram information captured by cumulative PCFG surprisal.

In the UCL corpus, accumulated PCFG surprisal (see Equation 4.4) did not improve reading time fit either ($p > 0.05$), unlike n-gram surprisal, which replicates the Dundee results. However, not even basic PCFG surprisal was predictive ($p > 0.05$) over this baseline model in the UCL corpus, whereas it was predictive over this baseline in the Dundee corpus. Posthoc testing on the exploratory data partition revealed that PCFG surprisal becomes predictive on the UCL corpus when the n-gram predictors are removed from the baseline ($p < 0.001$), which could indicate that PCFG surprisal simply helps predict reading times when the n-gram model is too weak. Alternatively, since UCL sentences were chosen for their brevity during corpus construction, there just may not be enough syntactic complexity in the corpus to provide an advantage to PCFG surprisal over the n-gram measures, which would explain why PCFG surprisal is still predictive for Dundee reading times where there is greater syntactic complexity.

4.8 Grammar Formalism Evaluation

One way of distinguishing the influence of syntactic PCFG surprisal compared to lexical n-gram surprisal is to look at long-distance dependencies, which cannot be reliably modeled with n-grams. This section uses the Dundee corpus to investigate whether a representation of hierarchical syntax that explicitly represents and preserves long-distance dependencies can improve reading time predictions over a hierarchic representation based on the Penn Treebank which discards long-distance dependencies. This evaluation compares total PCFG surprisal as calculated by the original Penn Treebank grammar to total PCFG surprisal calculated by the Nguyen et al. (2012) Generalized Categorial Grammar (GCG).

A GCG has a category set C, which consists of a set of primitive category types U, typically labeled with the part of speech of the head of a category (e.g. V, N, A, etc., for phrases or clauses headed by verbs, nouns, adjectives, etc.), followed by one or more unsatisfied dependencies, each consisting of an operator (-a and -b for adjacent argument dependencies preceding and following a head, -c and -d for adjacent conjunct dependencies preceding and following a head, -g for filler-gap dependencies, -r for relative pronoun dependencies, and some others), followed by a dependent category type. For example, the category for a transitive verb would be V-aN-bN, since it is headed by a verb and has unsatisfied dependencies to satisfied noun-headed categories preceding and following it (for the subject and direct object noun phrase, respectively).

As in other categorial grammars, inference rules for local argument attachment apply functors of category c-ad or c-bd to initial or final arguments of category d:

$$d \quad c\text{-a}d \Rightarrow c \tag{Aa}$$

$$c\text{-b}d \quad d \Rightarrow c \tag{Ab}$$

However, the Nguyen et al. (2012) GCG uses distinguished inference rules for modifier attachment, which allows modifier categories to be consolidated with categories for modifiers in other contexts (pre-verbal, post-verbal, etc.), and with certain predicative categories. This allows derivations in the training corpus involving different modifier types to also be consolidated, which increases the power of the extracted statistics. Inference rules for modifier attachment apply initial or final modifiers of category u-ad

to modificands of category c, for $u \in U$ and $c, d \in C$:

$$u\text{-a}d \quad c \Rightarrow c \tag{Ma}$$

$$c \quad u\text{-a}d \Rightarrow c \tag{Mb}$$

The Nguyen et al. (2012) GCG also uses distinguished inference rules to introduce, propagate, and bind missing non-local arguments, similar to the gap or slash rules of Generalized Phrase Structure Grammar (Gazdar et al., 1985) and Head-driven Phrase Structure Grammar (Pollard and Sag, 1994). Inference rules for gap attachment hypothesize gaps as initial arguments, final arguments, or modifiers, for $c, d \in C$:

$$c\text{-a}d \Rightarrow c\text{-g}d \tag{Ga}$$

$$c\text{-b}d \Rightarrow c\text{-g}d \tag{Gb}$$

$$c \Rightarrow c\text{-g}d \tag{Gc}$$

Non-local arguments, using non-local operator and argument category $\psi \in \{\text{-g}, \text{-h}, \text{-i}, \text{-r}\} \times C$, are then propagated to the consequent from all possible combinations of antecedents. For each rule $d \quad e \Rightarrow c \in \{\text{Aa–b,Ma–b}\}$:

$$d \quad e\psi \Rightarrow c\psi \tag{Ac–d,Mc–d}$$

$$d\psi \quad e \Rightarrow c\psi \tag{Ae–f,Me–f}$$

$$d\psi \quad e\psi \Rightarrow c\psi \tag{Ag–h,Mg–h}$$

In order to consolidate relative and interrogative pronouns in different pied-piping contexts into just two reusable categories, this grammar uses distinguished inference rules for relative and interrogative pronouns as well as tough constructions (e.g. this bread is easy to cut), which introduce clauses with gap dependencies, for $c, d, e \in C$,

$\psi \in \{\text{-g}\} \times C$:

$$d\text{-i}e \; c\text{-g}d \Rightarrow c\text{-i}e \tag{Fa}$$

$$d\text{-r}e \; c\text{-g}d \Rightarrow c\text{-r}e \tag{Fb}$$

$$c\text{-b}(d\psi) \; d\psi \Rightarrow c \tag{Fc}$$

Also, inference rules for relative pronoun attachment apply pronominal relative clauses of category c-rd to modificands of category e:

$$e \;\; c\text{-r}d \Rightarrow e \tag{R}$$

Because of its richer set of language-specific inference rules, the GCG grammar annotated by Nguyen et al. (2012) does not require different categories for words like which in different pied-piping contexts:

$$\dfrac{\dfrac{\text{cafes}}{\text{N}} \quad \dfrac{\dfrac{\text{which}}{\text{N-rN}} \quad \dfrac{\text{we ate in}}{\text{V-gN}}\,\text{Fb}}{\text{V-rN}}\,\text{R}}{\text{N}}$$

$$\dfrac{\dfrac{\text{cafes}}{\text{N}} \quad \dfrac{\dfrac{\dfrac{\text{in}}{\text{R-aN-bN}} \quad \dfrac{\text{which}}{\text{N-rN}}}{\text{R-aN-rN}}\,\text{Ab} \quad \dfrac{\dfrac{\text{we ate}}{\text{V}}}{\text{V-g(R-aN)}}\,\text{Gc}}{\text{V-rN}}\,\text{Fb}}{\text{N}}\,\text{R}$$

4.8.1 Experiment 3: Long-Distance Influences on Reading Times

Following van Schijndel et al. (2013b), the GCG calculation of PCFG surprisal comes from a GCG-reannotated version of the Penn Treebank whose grammar rules have undergone 3 iterations of the split-merge algorithm (Petrov et al., 2006). A k-best beam with a width of 5000 is used in order to be comparable to the PTB surprisal from the Section 4.7.

Significance testing is done as in the preceding evaluations: a baseline model is fit to reading times, each PCFG surprisal factor is added independently to the baseline,

Model	First Pass		Go-Past	
	Log-Likelihood	AIC	Log-Likelihood	AIC
Baseline	−1212242	2424592	−1261474	2523055
Base+PTB	−1212239*	2424587	−1261468*	2523047
Base+GCG	−1212239†	2424589	−1261470*	2523050
Base+Both	−1212235†	2424583	−1261465*	2523043

Baseline random slopes: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram, surp-GCG, surp-PTB; Baseline fixed effects: sentpos, wlen, rlen, prevfix, 5-gram, cumu-5-gram.

Table 4.6: Goodness of fit of models with differing syntactic calculations to reading times on the Dundee corpus. Significance testing was done between each model and the models in the section above it. Base+Both first pass significance applies to improvement over PTB ($p < .05$) and to improvement over GCG ($p < .01$), Base+Both go-past significance applies to improvement over each independent model. † $p < .05$ * $p < .01$

and both PCFG surprisal factors are added concurrently to the baseline. Each model is compared to the next simpler models using likelihood ratio tests.

The results (Table 4.6) show that GCG PCFG surprisal is a significant predictor of reading times even in the presence of the stronger n-gram baseline. Moreover, both PTB and GCG PCFG surprisal significantly improve reading time predictions even when the other PCFG surprisal measure is also included. This suggests that each is contributing something the other is not. Since the GCG grammar is derived from an automatically reannotated version of the Penn Treebank, there may be errors in the GCG annotation which cause errors in the estimates of underlying GCG structure. Since the PTB grammar is manually annotated by experts, the PTB grammar may be receiving credit for correct structural prediction in cases where GCG's estimates are incorrect. However, it seems likely that GCG may be providing a better fit in cases

of long-distance dependencies because such relations are not explicitly represented by the PTB grammar.

A follow-up evaluation using the experimental design from Section 4.7 but using GCG PCFG surprisal rather than PTB PCFG surprisal revealed that cumulative PCFG surprisal is still not a significant predictor when calculated using GCG. The failure of cumulative PCFG surprisal to improve over basic GCG PCFG surprisal could be expected since a strength of GCG is in enabling non-local decisions on a local basis (by propagating non-local decisions into the category labels), so any non-local advantage cumulative PCFG surprisal might confer is already compressed into the GCG categories.

The results of this evaluation suggest that reading times are mostly affected by local hierarchic structure, but the fact that GCG PCFG surprisal is able to provide a significant fit even in the presence of the PTB PCFG surprisal predictor suggests that some non-local information affects reading times. In particular, while this evaluation showed that accumulated syntactic context is not always a good predictor of reading times, some or all of the non-local information contained in the GCG categories may be used by readers and so influences reading time durations over the local structural information reflected in the PTB PCFG surprisal measure.

4.9 Discussion

The finding that hierarchic grammars orthogonally improve reading time predictions in each other's presence suggests that hierarchic structural information has a significant influence on reading times. Since both the PTB and GCG calculations of surprisal contain sequential information (e.g., of part-of-speech tags), if the effect in

this study was driven by purely sequential information as suggested by Frank and Bod (2011), one might expect either the PTB or the GCG calculations of surprisal (but not both) to be a significant predictor of reading times.

Instead, the present set of results support the hypothesis in Chapter 2 that non-local subcategorization decisions may have a strong influence on the (relatively early) reading time measures used in the present study. Such decisions would have to be conditioned on hierarchic structural information not present in either PTB PCFG surprisal or the sequential structure models of Frank and Bod (2011).

Further, predictability has been shown to affect word duration during speech production (Jurafsky et al., 2001; Aylett and Turk, 2006), and Demberg et al. (2012) found that hierarchic structure significantly improves over n-gram computations of predictability in that domain as well. Together, these findings suggest that hierarchic structure is not only a convenient descriptive tool for linguists, but that such structure is deeply rooted in the human language processor and is used during online language processing. However, the fact that PCFG surprisal is not improved via accumulation and its lack of predictivity in the UCL corpus suggests PCFG surprisal may only indirectly capture hierarchical syntactic influences and may need to be reformulated in order to best account for those effects.

Previous work has made a distinction between lexical surprisal, syntactic surprisal, and total surprisal (Demberg and Keller, 2008; Roark et al., 2009) as defined in the previous chapter. Fossum and Levy (2012) show that, with a non-cumulative bigram baseline, this distinction is not significant when predicting reading times, so the present study simply uses total surprisal. It may be interesting in future work

to see if the distinction between surprisal types becomes more or less useful as the sequential baseline improves.

The finding that cumulative n-gram information is useful in predicting reading times bears some resemblance to the finding that the spillover effect of a word is proportional to its logarithmic probability given the context (Smith and Levy, 2013). However, the spillover effect studied by Smith and Levy (2013) is one of a given fixation on the following fixation. Cumulative n-grams could conceivably capture spillover processing of parafovial preview obtained during the previous fixation, but generally cumulative n-grams permit finer predictability of a word given the unfixated intervening context, and outside of parafoveal preview, the two effects are quite different. Spillover is conceived as continued processing from the preceding fixation, while cumulative n-grams reflect the amount of new information contained in the region between one fixation and the next. Since the cumulative n-gram measure improves the predictability estimate of a word, it could provide a better measure of the spillover effect a given word will have on later fixations. Future work could investigate this by using the cumulative n-gram of a word to compute the predictability of the current word and the cumulative n-gram of the preceding region's word to predict the spillover effect from the preceding fixation. The present work suggests that doing so would provide even better reading time predictors.

4.10 Conclusion

First, this work suggests that the standard accounting for n-gram frequencies needs to change in psycholinguistic studies. Currently, the standard procedure is to use n-gram statistics only from the end of an eye-tracking region. This standard

calculates the influence of the final word in each region given the lexical context, but that context is never accounted for in regions greater than one word in length. Instead, psycholinguistic models need to additionally account for the probability of the context given its own preceding context to provide a coherent model of the probability of the observed lexical sequence.

The Dundee results in this chapter also suggest that, even with good cumulative and non-cumulative estimates of the frequency effects generated by a given lexical sequence, measures of hierarchic structure can provide a significant improvement to reading time predictions. However, the failure of accumulation to improve PCFG surprisal in both corpora, and the general lack of correlation between PCFG surprisal and reading times in the UCL corpus, suggests that PCFG surprisal may not be the best approximation of the syntactic frequency influences on human reading times, and other measures, formalisms and architectures for computing syntactic probability should be explored in the future.

Chapter 5: The influence of uncertainty on language processing*

This chapter shows that, although the correlation between PCFG surprisal and reading times is reduced after the n-gram improvements in the previous chapter, (un)certainty about upcoming hierarchical syntactic constructions (entropy) can reliably predict reading times even once upcoming lexical information is controlled for. This uncertainty has been estimated in previous studies by summing the log probabilities over hypotheses of possible sentence continuations, a process which is extremely expensive and so cannot be done with fine-grained grammars. This chapter shows that it is possible to estimate upcoming uncertainty by sampling the conditional probability distribution over future events. When the sample consists of events that actually occur, this estimate is equivalent to the surprisal of the upcoming events, which makes entropy estimation much less expensive to compute.

5.1 Introduction

The lexical frequencies of upcoming words affects reading times even when the upcoming word is masked from readers (Angele et al., 2015). Angele et al. suggest that the driving factor behind their result may be anticipation of upcoming difficulty.

*Portions of this chapter were originally published in van Schijndel and Schuler (2016) and others were published in van Schijndel and Schuler (2017).

For example, a less constraining context (i.e. less predictable upcoming words) may produce slower reading. This study uses information-theoretic entropy to test their hypothesis and to investigate the level of linguistic detail predicted by readers.

This work is scientifically important because it uses a large self-paced reading corpus to show that reading times are influenced both by uncertainty over upcoming syntactic constructions and by uncertainty over upcoming lexical items, which supports the hypothesis of Angele et al. (2015) that anticipation of upcoming difficulty influences reading times. While previous work has found evidence of prediction during language processing through responses to violated predictions (Wicha et al., 2004; Van Berkum et al., 2005; Fine et al., 2013; DeLong et al., 2014), the present work demonstrates that the influence of prediction can be reliably detected in reading times prior to any violation of that prediction. Other work, for example using a visual world paradigm (Altmann and Kamide, 1999; Kamide et al., 2003a; Ito and Speer, 2008), has also demonstrated predictive processing absent a prediction violation, but the present work demonstrates that such an effect is also observable in a broad-coverage self-paced reading corpus such as can be collected via Mechanical Turk. Finally, Roark et al. (2009) have previously shown that the entropy of upcoming syntactic categories influences self-paced reading times, but their entropy measure is extremely expensive to compute, they used a much smaller corpus,[27] and they did not find an influence of upcoming lexical uncertainty on reading times, unlike the present work.

In addition, this work demonstrates that surprisal (Hale, 2001; Levy, 2008), typically only used to estimate responses to observed stimuli, can be used to quantify predictive influences as well. From a computational perspective, this work provides an

[27]The corpus in this work is about 25 times larger.

inexpensive way to estimate the uncertainty experienced by readers, which will allow future studies to test the cognitive plausibility of various grammars and parsing algorithms, providing a tool with which to probe predictive human sentence processing outside of highly constraining experimental stimuli.

5.2 Background

Angele et al. (2015) tested whether lexical successor effects (influences of upcoming material) can be elicited even when readers were unable to view the upcoming words. They use a moving mask to hide upcoming words from readers but still find that the trigram predictability of the next hidden word is a significant predictor of reading times. Angele et al. (2015) hypothesize that readers may anticipate upcoming difficulty and slow down. That is, an unconstrained context with several plausible continuations might produce slower reading (due to each continuation's low predictability) than a highly constraining context with a smaller number of plausible continuations. To test this hypothesis, this work uses information-theoretic entropy to predict reading times.

Under information theory (Shannon, 1948), the entropy (H) of a random variable (X) is defined by the component probabilities of each possible value (x) of that variable:

$$H(X) = -\sum_{x \in X} \mathrm{P}(x) \log \mathrm{P}(x) \tag{5.1}$$

In the case of language processing, the possible values are words that have yet to be observed, and entropy is typically computed from the conditional probability of each possible value given the observations that have already been made.

Linzen and Jaeger (2015) distinguish single-step predictive entropy (uncertainty about the next processing step) from full entropy (uncertainty about the rest of the sentence). Since Angele et al. (2015) find that lexical frequency successor effects were only dependent on the word following a fixation, the present work is concerned with single-step predictive entropy. Linzen and Jaeger (2015) find that when single-step predictive entropy was computed over upcoming syntactic constituents based on verb subcategorization biases, it is not predictive of self-paced reading times. However, they hypothesize that the fit of entropy may improve when computed over finer-grained categories (they only compute probabilities for 6 subcategorization classes). The results in Analysis 4 of this chapter support their hypothesis.

Roark et al. (2009) define two variants of single-step predictive entropy to distinguish syntactic uncertainty from lexical uncertainty. Syntactic entropy is computed over the conditional probability of each preterminal (p) in the grammar (G) given the previously observed lexical sequence ($w_{1..i-1}$):

$$\mathrm{Syn}H^1_G(w_{1..i-1}) \stackrel{def}{=} \\ -\sum_{p_i \in G} \mathrm{P}_G(p_i \mid w_{1..i-1}) \log \mathrm{P}_G(p_i \mid w_{1..i-1}) \tag{5.2}$$

Syntactic entropy is computed in practice by generating all possible syntactic derivations[28] that can generate each possible upcoming word (w_i) in the vocabulary (V) and then subtracting from each derivation's log probability the emission probability of generating w_i from the chosen preterminal (p_i).

[28]In fact, the number of possible syntactic derivations is constrained by a very large beam.

Lexical entropy is computed over the conditional probability of each possible upcoming lexeme, given the previously observed lexical sequence:

$$\text{Lex}H^1_G(w_{1..i-1}) \stackrel{def}{=}$$
$$-\sum_{w_i \in V} \text{P}_G(w_i \mid w_{1..i-1}) \log \text{P}_G(w_i \mid w_{1..i-1}) \tag{5.3}$$

Roark et al. (2009) find that syntactic entropy is predictive of self-paced reading times but that lexical entropy is not, which this study is able to replicate on the Natural Stories corpus as well. Roark et al. suggest that the failure of lexical entropy to predict reading times may be due to the fact that their grammar is trained on the relatively small Brown portion of the Penn Treebank (Marcus et al., 1993), so their lexical probabilities may not be robust enough.

It is interesting to note that 'single-step prediction' was defined slightly differently for these two sets of authors. Roark et al. (2009) define it as a prediction over the next word in a lexical sequence, while Linzen and Jaeger (2015) define it as a prediction over the next syntactic category (e.g., noun phrase) that will branch from a partial derivation ending in a verb phrase. To avoid making a commitment as to the particular parsing strategy adopted by readers, this chapter will use the definition of 'single-step prediction' from Roark et al. (2009) to mean uncertainty about the next lexical observation.

5.3 Data

This study makes use of the Natural Stories self-paced reading corpus (Futrell et al., in prep). The corpus is a set of 10 texts (485 sentences) written to sound fluent and presented to subjects in order (unlike many constructed stimuli), but still containing many low-frequency and marked syntactic constructions, especially subject-

and object-extracted relative clauses, clefts, topicalized structures, extraposed relative clauses, sentential subjects, sentential complements, local structural ambiguity, and idioms. Self-paced reading time data was collected over these texts from 181 native English speakers. Reading times were excluded if they occurred at the beginning or end of a sentence, or if they were less than 100 ms or greater than 3000 ms. Approximately one third of the sentences (255,554 events) were used for exploration and two thirds of the sentences (512,469 events) were used as a confirmatory partition for significance testing to reduce the risk of false positives due to multiple comparisons.

5.4 Models

This study fits reading times using linear mixed effects models computed with the lme4 (version 1.1-7) R package (Bates et al., 2014). All models include a baseline of fixed effect predictors for word length, sentence position, and 5-gram surprisal. The models also include random intercepts for each word, each subject, and each subject/sentence pair. The last random intercept controls for the fact that multiple non-independent observations are drawn from each sentence. Finally, each model includes by-subject random slopes for all the fixed effects. All predictors were z-transformed prior to fitting. Significance values for each predictor were obtained using a likelihood ratio test between two mixed models: one of which contained both a by-subject random slope and a fixed effect for the predictor of interest (as well as all baseline predictors), and the other of which omitted the fixed effect for that predictor.

5.5 Analyses

5.5.1 Analysis 1: Single-Step Predictive Entropy

First, this work tries to replicate the findings of Roark et al. (2009) that single-step predictive entropy is predictive of reading times in this corpus as well. Single-step predictive syntactic and lexical entropy is computed here using the Roark (2001) top-down incremental parser. In keeping with the findings of Roark et al. (2009), syntactic entropy has a significant positive correlation with self-paced reading times in the Natural Stories corpus over the baseline model ($\hat{\beta} = 3.32$, $\hat{\sigma} = 0.52$, t-value $= 6.4$, p-value < 0.001). In contrast to Roark et al. (2009), an unusual correlation was observed between lexical entropy and reading times ($\hat{\beta} = -1.05$ ms; $\hat{\sigma} = 0.41$, t-value $= -2.5$, p-value $= 0.007$), though the effect does not remain significant after correcting for multiple comparisons. The lexical entropy effect seems especially unreliable since the effect is much smaller than that of syntactic entropy and goes in the opposite direction, and no such effect was observed either on the exploratory partition of the Natural Stories corpus[29] or by Roark et al. (2009). Further, the effect is no longer significant after correcting for multiple comparisons.

As Roark et al. (2009) point out, the unreliability of lexical entropy may stem from the sparseness of the training data. Unfortunately, computing predictive entropy is very expensive since it requires predictively running the parser over a large set of hallucinated observations whose cardinality is the size of the vocabulary for each

[29]Lexical entropy was used as a predictor on the confirmatory partition despite its lack of significance on the exploratory partition due to its direct relationship to the strong surprisal predictors in the next analyses. Still, the fact that lexical entropy obtained marginal significance with an unexpected sign on one set of reading times after failing to achieve significance on two other sets of reading times suggests the measure is not very reliable, though this could suggest more power is needed to observe any uncertainty effect with this entropy calculation.

actual observation. Therefore, meaningfully increasing the vocabulary is not generally practical.[30] The next section explores whether lexical entropy can be approximated effectively by another information-theoretic measure which is much easier to compute: surprisal.

5.5.2 Analysis 2: Surprisal as Entropy Approximation

Having shown that single-step predictive entropy is predictive for this data, this work now tests whether the surprisal of upcoming observations is an effective approximation of the influence of uncertaintyon reading times. Angele et al. (2015) found that trigram surprisal of an upcoming word is predictive of reading times and speculated that such an effect could be driven by uncertainty over future events, so this section tests whether the predictive entropy effect observed in Analysis 1 can be approximated by the PCFG surprisal of the upcoming word.

Roark (2011) showed that single-step predictive lexical entropy is equivalent to the expected value of total surprisal S:

$$S_G(w_i, w_{1..i-1}) \stackrel{def}{=} -\log \mathrm{P}_G(w_i \mid w_{1..i-1}) \tag{5.4}$$

$$\begin{aligned}
&\mathrm{Lex}H^1_G(w_{1..i-1}) \\
&\quad \stackrel{def}{=} \sum_{w_i \in V} -\mathrm{P}_G(w_i \mid w_{1..i-1}) \log \mathrm{P}_G(w_i \mid w_{1..i-1}) && (5.5)\\
&\quad = \sum_{w_i \in V} \mathrm{P}_G(w_i \mid w_{1..i-1})\ S_G(w_i, w_{1..i-1}) && (5.6)\\
&\quad = \mathrm{E}[S_G(w_i, w_{1..i-1})] && (5.7)
\end{aligned}$$

[30] An alternative to the approach taken in this chapter would be to maintain a constant vocabulary size but to train the conditional probabilities of that vocabulary over a much larger training set. Such an approach would only help if the weakness of lexical entropy is due to poor probability estimates rather than to unknown words.

where w_i is the current lexical item, $w_{1..i-1}$ is the sequence of previously observed lexical items and V is the vocabulary of the language. Therefore, total surprisal is a single sample from the conditional probability distribution over which single-step lexical entropy is computed, where the sampled observation is the occurrence that ultimately will be observed. Over several trials, then, future surprisal should approximate entropy since each observed occurrence should happen proportionately to its expected occurrence frequency. Further, future surprisal would be much cheaper to compute than single-step predictive lexical entropy because future surprisal omits the summation over all words in the vocabulary.

Since the Natural Stories corpus is a moving window self-paced reading corpus where even the segmentation of upcoming words was hidden from participants, any future surprisal effect observed in the current study cannot be explained as a symptom of parafoveal-on-foveal processing because participants were physically unable to see upcoming words, similar to the work of Angele et al. (2015).

To test total surprisal as an approximation of entropy, this work again uses the Roark (2001) parser and uses the total surprisal of each observation to predict the reading time of the preceding observation. This measure (future surprisal) also has significant positive correlations to reading times on both the exploratory and confirmatory partitions ($\hat{\beta} = 4.96$ ms, $\hat{\sigma} = 0.63$, t-value $= 7.9$, p-value < 0.001). This measure may be thought of as an aggregate approximation to entropy, whereas the lexical entropy output by the Roark (2001) parser may be thought of as a pointwise approximation to entropy. That is, Roark lexical entropy approximates the true lexical entropy for each new observation as the weighted average of the conditional

log-probability distribution over all possible upcoming observations at that point according to the parser's grammar. Future surprisal, in contrast, approximates the true lexical entropy when aggregated over the entire corpus by sampling from the conditional probability distribution the log-probability of each single observation that actually occurred. The fact that future surprisal is able to fit reading times more consistently (over both the exploratory and confirmatory partitions) than point-wise lexical entropy, but with a similar effect size, gives hope that this aggregate approach to entropy calculation is a cheaper and more robust means of computing entropy than a point-wise approximation.

5.5.3 Analysis 3: N-grams as Better Entropy Approximation

Although this work has shown that future surprisal can approximate the influence of uncertainty over reading times, it did so with a parser that was trained on a relatively small corpus using the coarse-grained Penn Treebank grammar. This section tests whether the approximation improves when trained on more data. In order to obtain conditional probabilities based on large amounts of data, this work uses a 5-gram back-off model computed with the KenLM toolkit (Heafield et al., 2013) on the Gigaword 4.0 corpus (Graff and Cieri, 2003), which consists of 2.96 billion words from English newswire text. Again, the 5-gram surprisal of each word was used to predict the reading times on the preceding word. Similar to future Roark surprisal (that is, surprisal computed with the Roark (2001) parser), future 5-gram surprisal has a significant positive correlation to reading times ($\hat{\beta} = 4.49$ ms, $\hat{\sigma} = 0.57$, t-value 7.9, p-value < 0.001), and when future 5-gram surprisal is in the model, future Roark surprisal ceases to be a significant predictor of reading times.

Although n-gram surprisal seems to account for the effect of Roark lexical entropy, with an even larger effect size,[31] such a measure is unable to account theoretically for the effect of Roark syntactic entropy, which omits lexical emission probabilities. The next section explores whether humans predict upcoming material with greater syntactic specificity than the tag set provided with the Penn Treebank insofar as such prediction affects their self-paced reading times.

5.5.4 Analysis 4: Fine-Grained Syntactic Prediction

Although future n-gram surprisal seems to account for a lexical entropy effect, it is unable to account theoretically for the effect of Roark syntactic entropy, since n-gram surprisal reflects lexical probabilities and syntactic entropy reflects syntactic probabilities (without lexical emission probabilities). However, future Roark PCFG surprisal using the default set of Penn Treebank syntactic categories was unable to predict reading times when future n-gram surprisal was in the model. Previous work on predictive processing has suggested that predictions can be relatively fine-grained (Luke and Christiansen, 2015; Kim and Lai, 2012), so this section explores whether humans predict upcoming material with fine-grained syntactic specificity.

Whereas the above analyses used the Roark (2001) parser with the default Penn Treebank tag set, this section uses the van Schijndel et al. (2013a) parser, which computes surprisal using the Petrov et al. (2006) latent-variable grammar computed from sections 2-21 of the Wall Street Journal portion of the Penn Treebank and thereby achieves higher parsing accuracy than the Roark parser (see Chapter 3). The latent-variable grammar is derived from a split-merge algorithm that creates

[31] Readers may be concerned that the effect size difference could have been produced by different predictor ranges, but the predictors have similar ranges since both were z-transformed prior to fitting the models.

fine-grained subcategory tags from the basic Penn Treebank category tags. For this analysis, the grammar underwent 5 split-merge operations to obtain optimally tuned tags, following the recommendations of Petrov et al.

Unlike future Roark surprisal, future latent-variable surprisal is able to obtain a significant positive correlation with reading times, even in the presence of both future-5-gram surprisal and Roark syntactic entropy ($\hat{\beta} = 4.10$ ms, $\hat{\sigma} = 0.74$, t-value $= 5.6$, p-value < 0.001). Similarly, Roark syntactic entropy retains its predictivity even in the presence of future 5-gram surprisal and future latent-variable surprisal ($\hat{\beta} = 4.62$ ms, $\hat{\sigma} = 0.52$, t-value $= 8.8$, p-value < 0.001), suggesting that there is still an advantage to approximating syntactic entropy using the costly pointwise method. This may be further evidence that PCFG surprisal is not the best approximation of human syntactic frequency expectations, though it seems sufficient to capture at least some of the influence of syntactic uncertainty on self-paced reading times.

5.5.5 Analysis 5: Future Surprisal for Eye-Tracking

Having shown in the previous sections the effectiveness of future surprisal at predicting self-paced reading times, this analysis uses the future 5-gram surprisal and future latent variable PCFG surprisal measures to predict first pass reading times on the Dundee (Kennedy et al., 2003) and UCL (Frank et al., 2015) eye-tracking corpora. As in the previous chapter, the first and last fixation of each sentence and line was omitted from analysis, as were fixations whose corresponding region length was greater than 4. Because this analysis is conducted over eye-tracking data, and because the previous chapter demonstrated that accumulation can improve the fit of

surprisal to reading times in eye-tracking over non-cumulative surprisal, this analysis uses future cumulative surprisal measures rather than the future non-cumulative measures used in the self-paced reading analyses above.[32] The predictors of interest were tested over a baseline model that included fixed effects and by-subject random slopes for sentence position, word length, region length, whether or not the previous word was fixated, cumulative 5-gram surprisal, and the future region length. In addition, the models included random intercepts for subjects, items, and subject/sentence pairings.

On the UCL corpus, future cumulative 5-gram surprisal is predictive over the baseline model ($\hat{\beta} = 4.7$, $p < 0.001$). Future cumulative PCFG surprisal, however, is not a significant predictor of UCL reading times ($p = 0.07$) unlike in the self-paced reading setting. Neither predictor is significant in the Dundee corpus (5-grams: $p = 0.24$; PCFG: $p = 0.73$), but it may be that any effect is hidden by the smaller number of subjects or that the lack of detailed comprehension questions caused subjects to forgo systematic predictive processing. The UCL results reinforce the previous findings that future cumulative 5-gram surprisal provides a good estimate of uncertainty over upcoming material in eye tracking as well as in self-paced reading, though the UCL sentences may be syntactically too simple to provide a good testbed for syntactic prediction effects.

5.5.6 Analysis 6: Limitations of successor n-grams

Since this work has shown the effectiveness of future n-gram surprisal at predicting reading times in both self-paced and eye-tracking settings, this section explores the

[32]Because accumulation did not seem to benefit PCFG surprisal in the previous chapter, future work could verify whether future non-cumulative PCFG surprisal obtains a better fit to reading times than future cumulative PCFG surprisal.

extent of the prediction in both reading time paradigms in the UCL and Natural Stories corpora. On the exploration partitions, four cumulative 5-gram successor predictors are tested which utilize look-ahead for 1-word, 2-words, 3-words, or 4-words. Each future n-gram variant is a forward 5-gram measure that accumulates over the given number of successor words. In the UCL eye-tracking setting, each measure only includes material up to the following fixation, so 4-word future n-grams compute future cumulative n-gram probabilities up to four words ahead, but if the upcoming saccade is only two words long, then 4-word future n-grams will only compute future n-gram probability for the upcoming two words. In the Natural Stories self-paced reading setting, each measure accumulates over increasing amounts of material after each word, so 4-word future n-grams compute future cumulative 5-gram probabilities for the 4 words after the observed word. Each future n-gram variant is evaluated based on how it improves over a baseline mixed model. The Natural Stories baseline contains fixed effects and by-subject random slopes for sentence position, word length, 5-gram surprisal, and total surprisal. The UCL baseline contains fixed effects and by-subject random slopes for sentence position, word length, region length, whether the previous word was fixated, and cumulative 5-gram surprisal.

In the UCL corpus, 2-word future n-grams provide the best fit to the data ($p < 0.001$) even though there are 3- and 4-word saccades in the data. In the Natural Stories corpus, 1-word future n-grams provide the best fit to the data ($p < 0.001$), which is in line with the findings of Angele et al. (2015). The fact that the self-paced reading results align with the masked eye-tracking results of Angele et al. (2015), while the unmasked eye-tracking results show a slightly larger predicted distance, suggests that the successor effect observed by Angele et al. may only account for a subset of the

successor influences on reading times. It's possible that parafoveal preview, which was not possible in the masked condition of the Angele et al. (2015) study or in the Natural Stories corpus, accounts for the additional look-ahead observed in the UCL corpus (e.g., parafoveal look-ahead could help with the word following the target, and the predictive effect observed by Angele et al. could help with the next word), but additional investigation of this hypothesis is left for future work.

5.6 Discussion

While previous work has found evidence of prediction during language processing through responses to violated predictions (Wicha et al., 2004; Van Berkum et al., 2005; Fine et al., 2013; DeLong et al., 2014), the present work demonstrates that the influence of prediction can be reliably detected in reading times prior to any violation of that prediction. Other work, for example using a visual world paradigm (Altmann and Kamide, 1999; Kamide et al., 2003a; Ito and Speer, 2008), has also demonstrated predictive processing absent a prediction violation, but the present work demonstrates that such an effect is also observable in a broad-coverage self-paced reading corpus such as can be collected via Mechanical Turk.

Previous studies have claimed that a positive correlation between entropy and reading times would indicate that there is a competition cost between multiple parse hypotheses (Linzen and Jaeger, 2015), but this is not the only possible explanation for such a correlation. Instead, it seems reasonable that if readers have more uncertainty about upcoming material, they would slow their reading in order to better process the less expected information (reducing their expected per-millisecond surprise). If,

instead, readers are reasonably confident about what words they are about to encounter, they may speed up in order to maximize the per millisecond informativity of their observations. One extreme example of this type of speed tuning may be seen with determiners and other function words which readers are likely to skip entirely in eye-tracking analyses. This sort of tuning may be exaggerated in the moving window self-paced reading paradigm, where readers will be unable to regress if they speed past an unexpected observation, which could be why this effect shows up more strongly in self-paced reading.

The finding that future latent-variable surprisal can fit self-paced reading times even in the presence of future 5-gram surprisal suggests that humans make syntactic predictions as well as lexical predictions. And the fact that this predictivity holds even in the presence of Roark syntactic entropy suggests that sampling the conditional probability distribution with future surprisal is a robust (though possibly incomplete) means of approximating entropy. Surprisal is computationally much cheaper than entropy, and it can therefore provide samples from a much more fine-grained conditional probability distribution over possible analyses than would be practical for entropy calculation. The finding that future total Roark surprisal cannot predict reading times in the presence of good n-gram models suggests that humans predict upcoming syntactic material at a very fine-grained level. This interpretation is also supported by the fact that Linzen and Jaeger (2015) found that a very coarse entropy, computed over subcategorization frequencies over six types of syntactic constituents, was not predictive of reading times.

The finding that Roark syntactic entropy retains its reading time predictivity in the presence of future 5-gram surprisal and future latent-variable surprisal may

suggest that humans estimate their certainty about upcoming parses in a point-wise manner, meaning that they likely evaluate the likelihood of multiple parse hypotheses in parallel. Such a finding is consistent with parallel models of sentence processing but may be problematic for serial processing models. Another interpretation of this finding is that a point-wise entropy approximation is more stable and so can serve as a back-off for the less stable but more nuanced aggregate approximations provided by both the n-gram and latent-variable surprisal models. It is left to future work to differentiate between these two possibilities.

It may seem strange that total latent-variable surprisal was used in this study instead of syntactic latent-variable surprisal since the goal of moving beyond future n-gram surprisal was to capture something of syntactic entropy, which omits lexical emission probabilities; however, explorations on the development partition revealed that future total surprisal generally provides better fits to reading times than future syntactic surprisal even in the presence of future 5-gram surprisal. In any case, the goal was not necessarily to approximate Roark syntactic entropy but to capture an aspect of the uncertainty experienced by readers, of which Roark lexical entropy and Roark syntactic entropy are themselves approximations. In fact, the consistent correlation between future surprisal (both n-gram and latent-variable) and reading times compared to Roark lexical entropy suggests that fine-grained aggregate entropy approximation via future surprisal is more robust than the coarser but more intuitive point-wise lexical entropy approximation output by the Roark (2001) parser.

It is important to keep in mind that the entropy findings in this chapter are distinct from those in the entropy reduction literature. The analyses in this chapter highlight a broad-coverage correlation of fine-grained predictive entropy to self-paced

reading times and show that people slow down before areas of greater uncertainty, though they may also slow down due to larger information gains. These effects are not necessarily mutually exclusive because entropy reduction (ΔH) deals with changes in entropy while predictive entropy deals with the overall level of uncertainty in a text. That is, an entropy reduction of k may produce the same $k \cdot \beta_{\Delta H}$ ms effect on reading times whether the resulting entropy is low or high. However, the results in this chapter do have implications for entropy reduction since they suggest humans are sensitive to fine-grained entropy, which should be considered by psycholinguists seeking to estimate the influence of entropy reduction.

5.7 Conclusion

This chapter has replicated previous findings that single-step predictive entropy is positively correlated with reading times and presented new results that show this correlation can be cheaply approximated using both future n-gram surprisal and future latent-variable PCFG surprisal. The present results also demonstrate that approximations of entropy improve prediction of self-paced reading times as the granularity of the approximation increases. The findings in this chapter support the hypothesis of Angele et al. (2015) that single-step predictive entropy is correlated with reading times, which suggests that at least some effects that have previously been attributed to parafoveal-on-foveal influences (which cannot be present in moving window self-paced reading) may be due to effects of anticipated uncertainty.

Chapter 6: Long-distance syntactic dependencies in acquisition*

Hierarchic syntax affects comprehension and production, but understandable grammar must be acquired somehow. Many previous studies have outlined syntactic acquisition models, but generally these models are built for engineering purposes (e.g., rapid deployment in low-resource languages) rather than to model the time course of linguistic development. In contrast, this chapter focuses on modeling a set of developmental phenomena that have been largely neglected by previous computational models. In particular, this chapter explores how long-distance dependencies can be acquired by human English learners at an extremely young age even in the absence of explicit cues. Developing an acquisition model that closely replicates the linguistic development of children could provide insight into which frequencies are being tracked by humans, which could help build better contextual dependencies into complexity measures such as syntactic surprisal.

6.1 Introduction

The phenomenon of filler-gap, where the argument of a predicate appears outside its canonical position in the phrase structure (e.g. [the apple]$_i$ that the boy ate

*The work in this chapter was originally published in van Schijndel and Elsner (2014).

t_i was good. or [What]$_i$ did the boy eat t_i?), has long been an object of study for syntacticians (Ross, 1967) due to its apparent processing complexity. Such complexity is due, in part, to the arbitrary length of the dependency between a filler and its gap (e.g. [the apple]$_i$ that Mary said the boy ate t_i).

Recent studies indicate that comprehension of filler-gap constructions begins around 15 months (Seidl et al., 2003; Gagliardi et al., in prep). This finding raises the question of how such a complex phenomenon could be acquired so early since children at that age do not yet have a very advanced grasp of language (e.g. ditransitives do not seem to be generalized until at least 31 months; Goldberg et al. 2004, Bello 2012). This work shows that filler-gap comprehension in English may be acquired through learning word orderings rather than relying on hierarchical syntactic knowledge.

This work describes a cognitive model of the developmental timecourse of filler-gap comprehension with the goal of setting a lower bound on the modeling assumptions necessary for an ideal learner to display filler-gap comprehension. In particular, the model described in this paper takes chunked child-directed speech as input and learns orderings over semantic roles. These orderings then permit the model to successfully resolve filler-gap dependencies.[33] Further, the model presented here is also shown to initially reflect an idiosyncratic role assignment error observed in development (e.g. A and B kradded interpreted as A kradded B; Gertner and Fisher, 2012), though after training, the model is able to avoid the error. As such, this work may be said to model a learner from 15 months to between 25 and 30 months.

[33]This model does not explicitly learn gap positions, but rather assigns thematic roles to arguments based on where those arguments are expected to manifest. This approach to filler-gap comprehension is supported by findings that show people do not actually link fillers to gap positions but instead link the filler to a verb with missing arguments (Pickering and Barry, 1991)

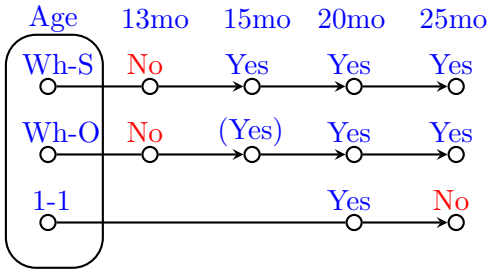

Figure 6.1: The developmental timeline of subject (Wh-S) and object (Wh-O) wh-clause extraction comprehension suggested by experimental results (Seidl et al., 2003; Gagliardi et al., in prep). Parentheses indicate weak comprehension. The final row shows the timeline of 1-1 role bias errors (Naigles, 1990; Gertner and Fisher, 2012). Missing nodes denote a lack of studies.

6.2 Background

The developmental timeline during which children acquire the ability to process filler-gap constructions is not well understood. Language comprehension precedes production, and the developmental literature on the acquisition of filler-gap constructions is sparsely populated due to difficulties in designing experiments to test filler-gap comprehension in preverbal infants. Older studies typically looked at verbal children and the mistakes they make to gain insight into the acquisition process (de Villiers and Roeper, 1995).

Recent studies, however, indicate that filler-gap comprehension likely begins earlier than production (Seidl et al., 2003; Gagliardi and Lidz, 2010; Gagliardi et al., in prep). Therefore, studies of verbal children are probably actually testing the acquisition of production mechanisms (planning, motor skills, greater facility with lexical access, etc) rather than the acquisition of filler-gap. Note that these may be related

since filler-gap could introduce greater processing load which could overwhelm the child's fragile production capacity (Phillips, 2010).

Seidl et al. (2003) showed that children are able to process wh-extractions from subject position (e.g. [Who]$_i$ t$_i$ ate pie?) as young as 15 months while similar extractions from object position (e.g. [What]$_i$ did the boy eat t$_i$?) remain unparseable until around 20 months of age.[34] This line of investigation has been reopened and expanded by Gagliardi et al. (in prep) whose results suggest that the experimental methodology employed by Seidl et al. (2003) was flawed in that it presumed infants have ideal performance mechanisms. By providing more trials of each condition and controlling for the pragmatic felicity of test statements, Gagliardi et al. (in prep) provide evidence that 15-month old infants can process wh-extractions from both subject and object positions. Object extractions are more difficult to comprehend than subject extractions, however, perhaps due to additional processing load in object extractions (Gibson, 1998; Phillips, 2010). Similarly, Gagliardi and Lidz (2010) show that relativized extractions with a wh-relativizer (e.g. Find [the boy]$_i$ who t$_i$ ate the apple.) are easier to comprehend than relativized extractions with that as the relativizer (e.g. Find [the boy]$_i$ that t$_i$ ate the apple.).

Yuan et al. (2012) demonstrate that 19-month olds use their knowledge of nouns to learn both verbs and their associated argument structure. In their study, infants were shown video of a person talking on a phone using a nonce verb with either one or two nouns (e.g. Mary kradded Susan). Under the assumption that infants look longer at things that correspond to their understanding of a prompt, the infants were

[34]Since the wh-phrase is in the same (or a very similar) position as the original subject when the wh-phrase takes subject position, it is not clear that these constructions are true extractions (Culicover, 2013), however, this paper will continue to refer to them as such for ease of exposition.

then shown two images that potentially depicted the described action – one picture where two actors acted independently (reflecting an intransitive proposition) and one picture where one actor acted on the other (reflecting a transitive proposition).[35] Even though the infants had no extralinguistic knowledge about the verb, they consistently treated the verb as transitive if two nouns were present and intransitive if only one noun was present.

Similarly, Gertner and Fisher (2012) show that intransitive phrases with conjoined subjects (e.g. John and Mary gorped) are given a transitive interpretation (i.e. John gorped Mary) at 21 months (henceforth termed '1-1 role bias'), though this effect is no longer present at 25 months (Naigles, 1990). This finding suggests both that learners will ignore canonical structure in favor of using all possible arguments and that children have a bias to assign a unique semantic role to each argument. It is important to note, however, that cross-linguistically children do not seem to generalize beyond two arguments until after at least 31 months of age (Goldberg et al., 2004; Bello, 2012), so a predicate occurring with three nouns would still likely be interpreted as merely transitive rather than ditransitive.

Computational modeling provides a way to test the computational level of processing (Marr, 1982). That is, given the input (child-directed speech, adult-directed speech, and environmental experiences), it is possible to probe the computational processes that result in the observed output. However, previous computational models of grammar induction (Klein and Manning, 2004), including infant grammar induction (Kwiatkowski et al., 2012), have not addressed filler-gap comprehension.[36]

[35]There were two actors in each image to avoid biasing the infants to look at the image with more actors.

[36]Joshi et al. (1990) and subsequent work show that filler-gap phenomena can be formally captured by mildly context-sensitive grammar formalisms; these have the virtue of scaling up to adult

Susan	said	John	gave	girl	book
-3	-2	-1	0	1	2

Table 6.1: An example of a chunked sentence (Susan said John gave the girl a red book) with the sentence positions labelled. Nominal heads of noun chunks are in bold.

The closest work to that presented here is the work on BabySRL (Connor et al., 2008, 2009, 2010). BabySRL is a computational model of semantic role acquistion using a similar set of assumptions to the current work. BabySRL learns weights over ordering constraints (e.g. preverbal, second noun, etc.) to acquire semantic role labelling while still exhibiting 1-1 role bias. However, no analysis has evaluated the ability of BabySRL to acquire filler-gap constructions. Further comparison to BabySRL may be found in Section 6.6.

6.3 Assumptions

The present work restricts itself to acquiring filler-gap comprehension in English. The model presented here learns a single, non-recursive ordering for the semantic roles in each sentence relative to the verb since several studies have suggested that early child grammars may consist of simple linear grammars that are dictated by semantic roles (Diessel and Tomasello, 2001; Jackendoff and Wittenberg, in press). This work assumes learners can already identify nouns and verbs, which is supported by Shi et al. (1999) who show that children at an extremely young age can distinguish between content and function words and by Waxman and Booth (2001) who show that children

grammar, but due to their complexity, do not seem to have been described as models of early acquisition.

can distinguish between different types of content words. Further, since Waxman and Booth (2001) demonstrate that, by 14 months, children are able to distinguish nouns from modifiers, this work assumes learners can already chunk nouns and access the nominal head. To handle recursion, this work assumes that children treat the final verb in each sentence as the main verb (implicitly assuming sentence segmentation), which ideally assigns roles to each of the nouns in the sentence.

Due to the findings of Yuan et al. (2012) that children use knowledge of nouns to acquire verbal argument structure, this work adopts a 'syntactic bootstrapping' theory of acquisition (Gleitman, 1990), where structural properties (e.g. number of nouns) inform the learner about semantic properties of a predicate (e.g. how many semantic roles it confers). Since infants infer the number of semantic roles, this work further assumes they already have expectations about where these roles tend to be realized in sentences, if they appear. These positions may correspond to different semantic roles for different predicates (e.g. the subject of run and of melt); however, the role for predicates with a single argument is usually assigned to the noun that precedes the verb while a second argument is usually assigned after the verb. The semantic properties of these roles may be learned lexically for each predicate, but that is beyond the scope of this work. Therefore, this work uses syntactic and semantic roles interchangeably (e.g. subject and agent).

Finally, following the finding by Gertner and Fisher (2012) that children interpret intransitives with conjoined subjects as transitives, this work assumes that semantic roles have a one-to-one correspondence with nouns in a sentence (similarly used as a soft constraint in the semantic role labelling work of Titov and Klementiev, 2012).

	μ	σ	π
G_{SC}	-1	0.5	.999
G_{SN}	-1	3	.001
G_{OC}	1	0.5	.999
G_{ON}	1	3	.001
Φ	.00001		

Table 6.2: Initial values for the mean (μ), standard deviation (σ), and prior (π) of each Gaussian as well as the skip penalty (Φ) used in this paper.

6.4 Model

The model represents the preferred locations of semantic roles relative to the verb as distributions over real numbers. This idea is adapted from Boersma (1997) who uses it to learn constraint rankings in optimality theory.

In this work, the final (main) verb is placed at position 0; words (and chunks) before the verb are given progressively more negative positions, and words after the verb are given progressively more positive positions (see Table 6.1). Learner expectations of where an argument will appear relative to the verb are modelled as two-component Gaussian mixtures: one mixture of Gaussians ($G_{S.}$) corresponds to the subject argument, another ($G_{O.}$) corresponds to the object argument. There is no mixture for a third argument since children do not generalize beyond two arguments until later in development (Goldberg et al., 2004; Bello, 2012).

One component of each mixture learns to represent the canonical position for the argument ($G_{.C}$) while the other ($G_{.N}$) represents some alternate, non-canonical position such as the filler position in filler-gap constructions. To reflect the fact that learners have had 15 months of exposure to their language before acquiring filler-gap,

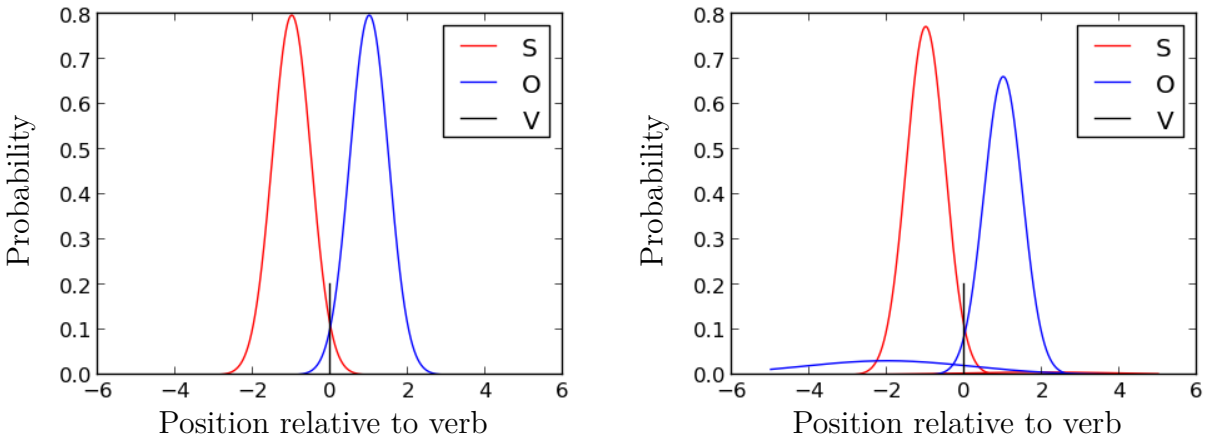

Figure 6.2: Visual representations of (Left) the initial model's expectations of where arguments will appear, given the initial parameters in Table 6.2 and (Right) the converged model's expectations of where arguments will appear.

the mixture is initialized so that there is a stronger probability associated with the canonical Gaussian than with the non-canonical Gaussian of each mixture.[37] Finally, the one-to-one role bias is explicitly encoded such that the model cannot use a label that has already been used elsewhere in the sentence.

Thus, the initial model conditions (see Figure 6.2) are most likely to realize an SVO ordering, although it is possible to obtain SOV (by sampling a negative number from the blue curve) or even OSV (by also sampling the red curve very close to 0). The model is most likely to hypothesize a preverbal object when it has already assigned the subject role to something and, in addition, there is no postverbal noun competing for the object label. In other words, the model infers that an object extraction may have occurred if there is a 'missing' postverbal argument.

[37]Akhtar (1999) finds that learners may not have strong expectations of canonical argument positions until four years of age, but the results of the current study are extremely robust to changes in initialization, as discussed in Section 6.7 of this paper, so this assumption is mostly adopted for ease of exposition.

Finally, the probability of a given sequence is the product of the label probabilities for the component argument positions (e.g. G_{SC} generating an argument at position -2, etc). Since many sentences have more than two nouns, the model is allowed to skip nouns by multiplying a penalty term (Φ) into the product for each skipped noun; the cost is set at 0.00001 for this study, though see Section 6.7 for a discussion of the constraints on this parameter. See Table 6.2 for initialization parameters and Figure 6.2 for a visual representation of the initial expectations of the model.

This work uses a model with 2-component mixtures for both subjects and objects (termed the symmetric model). This formulation achieves the best fit to the training data according to the Bayesian Information Criterion (BIC).[38] However, follow-up experiments find that the non-canonical subject Gaussian only improves the likelihood of the data by erroneously modeling postverbal nouns in imperative statements. The lack of a canonical subject in English imperatives allows the model to improve the likelihood of the data by using the non-canonical subject Gaussian to capture fictitious postverbal arguments. When imperatives are filtered out of the training corpus, the symmetric model obtains a worse BIC fit than a model that lacks the non-canonical subject Gaussian. Therefore, if one makes the assumption that imperatives are prosodically-marked for learners (e.g. the learner is the implicit subject), the best model is one that lacks a non-canonical subject.[39] The remainder of this paper assumes a symmetric model to demonstrate what happens if such an assumption is not made; for the evaluations described in this paper, the results are similar in either case.

[38]The BIC rewards improved log-likelihood but penalizes increased model complexity.

[39]This finding suggests that a Dirichlet Process or other means of dynamically determining the number of components in each mixture would converge to a model that lacks non-canonical subjects if imperative filtering were employed.

This model differs from other non-recursive computational models of grammar induction (e.g. Goldwater and Griffiths, 2007) since it is not based on Hidden Markov Models. Instead, it determines the best ordering for the sentence as a whole. This approach bears some similarity to a Generalized Mallows model (Chen et al., 2009), but the current formulation was chosen due to being independently posited as cognitively plausible (Boersma, 1997).

Figure 6.2 (Right) shows the converged, final state of the model. The model expects the first argument (usually agent) to be assigned preverbally and expects the second (say, patient) to be assigned postverbally; however, there is now a larger chance that the second argument will appear preverbally.

6.5 Evaluation

The model in this work is trained using transcribed child-directed speech (CDS) from the BabySRL portions (Connor et al., 2008) of CHILDES (MacWhinney, 2000). Chunking is performed using a basic noun-chunker from NLTK (Bird et al., 2009). Based on an initial analysis of chunker performance, yes is hand-corrected to not be a noun. Poor chunker perfomance is likely due to a mismatch in chunker training and testing domains (Wall Street Journal text vs transcribed speech), but chunking noise may be a good estimation of learner uncertainty, so the remaining text is left uncorrected. All noun phrase chunks are then replaced with their final noun (presumed the head) to approximate the ability of children to distinguish nouns from modifiers (Waxman and Booth, 2001). Finally, for each sentence, the model assigns sentence positions to each word with the final verb at zero.

Viterbi Expectation-Maximization is performed over each sentence in the corpus to infer the parameters of the model. During the Expectation step, the model uses the current Gaussian parameters to label the nouns in each sentence with argument roles. Since the model is not lexicalized, these roles correspond to the semantic roles most commonly associated with subject and object. The model then chooses the best label sequence for each sentence.

These newly labelled sentences are used during the Maximization step to determine the Gaussian parameters that maximize the likelihood of that labelling. The mean of each Gaussian is updated to the mean position of the words it labels. Similarly, the standard deviation of each Gaussian is updated with the standard deviation of the positions it labels. A learning rate of 0.3 is used to prevent large parameter jumps. The prior probability of each Gaussian is updated as the ratio of that Gaussian's labellings to the total number of labellings from that mixture in the corpus:

$$\pi_{\rho\theta} = \frac{\mid G_{\rho\theta} \mid}{\mid G_{\rho\cdot} \mid} \tag{6.1}$$

where $\rho \in \{S, O\}$ and $\theta \in \{C, N\}$.

Best results seem to be obtained when the skip-penalty is loosened by an order of magnitude during testing. Essentially, this forces the model to tightly adhere to the perceived argument structure during training to learn more rigid parameters, but the model is allowed more leeway to skip arguments it has less confidence in during testing. Convergence (see Figure 6.2) tends to occur after four iterations but can take up to ten iterations depending on the initial parameters.

	Eve (n = 4820)			Adam (n = 4461)		
	P	R	F	P	R	F
Initial	.54	.64	.59	.53	.60	.56
Trained	.52	.69	.59*	.51	.65	.57*
Initial$_c$	.56	.66	.60	.55	.62	.58
Trained$_c$	.54	.71	.61*	.53	.67	.59*

Table 6.3: Overall accuracy on the Eve and Adam sections of the BabySRL corpus. Bottom rows reflect accuracy when non-agent roles are collapsed into a single role. Note that improvements are numerically slight since filler-gap is relatively rare (Schuler, 2011). $^*p << .01$

Subject Extraction filter:	S	x	V	…
Object Extraction filter:	O	…	V	…

	Eve (n = 1345)			Adam (n = 1287)		
	P	R	F	P	R	F
Initial$_c$	.53	.57	.55	.53	.52	.52
Trained$_c$	.55	.67	.61*	.54	.63	.58*

Table 6.4: (Above) Filters to extract filler-gap constructions: A) the subject and verb are not adjacent, B) the object precedes the verb. (Below) Filler-gap accuracy on the Eve and Adam sections of the BabySRL corpus when non-agent roles are collapsed into a single role. $^*p << .01$

Since the model is unsupervised, it is trained on a given corpus (e.g. Eve) before being tested on the role annotations of that same corpus. The Eve corpus was used for development purposes,[40] and the Adam data was used only for testing.

For testing, this study uses the semantic role annotations in the BabySRL corpus. These annotations were obtained by automatically semantic role labelling portions of

[40]This is included for transparency, though the initial parameters have very little bearing on the final results as stated in Section 6.7, so the danger of overfitting to development data is very slight.

	P	R	F	P	R	F
Eve	Subj (n = 691)			Obj (n = 654)		
Initial$_c$	.66	.83	.74	.35	.31	.33
Trained$_c$	.64	.84	.72†	.45	.52	.48*
Adam	Subj (n = 886)			Obj (n = 1050)		
Initial$_c$	.69	.81	.74	.33	.27	.30
Trained$_c$	.66	.81	.73	.44	.48	.46*
	P	R	F	P	R	F
Eve	Wh- (n = 689)			That (n = 125)		
Initial$_c$	.63	.45	.53	.43	.48	.45
Trained$_c$	.73	.75	.74*	.44	.57	.50†
Adam	Wh- (n = 748)			That (n = 189)		
Initial$_c$	.50	.37	.42	.50	.50	.50
Trained$_c$	.61	.65	.63*	.47	.56	.51†

Table 6.5: (Left) Subject-extraction accuracy and object-extraction accuracy and (Right) Wh-relative accuracy and that-relative accuracy; calculated over the Eve and Adam sections of the BabySRL corpus with non-agent roles collapsed into a single role. $^\dagger p = .02$ $^* p << .01$

CHILDES with the system of Punyakanok et al. (2008) before roughly hand-correcting them (Connor et al., 2008). The BabySRL corpus is annotated with 5 different roles, but the model described in this paper only uses 2 roles. Therefore, overall accuracy results (see Table 6.3) are presented both for the raw BabySRL corpus and for a collapsed BabySRL corpus where all non-agent roles are collapsed into a single role (denoted by a subscript $_c$ in all tables).

Since children do not generalize above two arguments during the modelled age range (Goldberg et al., 2004; Bello, 2012), the collapsed numbers more closely reflect the performance of a learner at this age than the raw numbers. The increase in accuracy obtained from collapsing non-agent arguments indicates that children may

initially generalize incorrectly to some verbs and would need to learn lexically-specific role assignments (e.g. double-object constructions of give). Since the current work is interested in general filler-gap comprehension at this age, including over unknown verbs, the remaining analyses in this paper consider performance when non-agent arguments are collapsed.[41]

Next, a filler-gap version of the BabySRL corpus is created using a coarse filtering process: the new corpus is comprised of all sentences where an associated object precedes the final verb and all sentences where the relevant subject is not immediately followed by the final verb (see Table 6.4). For these filler-gap evaluations, the model is trained on the full version of the corpus in question (e.g. Eve) before being tested on the filler-gap subset of that corpus. The overall results of the filler-gap evaluation (see Table 6.4) indicate that the model improves significantly at parsing filler-gap constructions after training.

The performance of the model on role-assignment in filler-gap constructions may be analyzed further in terms of how the model performs on subject-extractions compared with object-extractions and in terms of how the model performs on that-relatives compared with wh-relatives (see Table 6.5).

The model actually performs worse at subject-extractions after training than before training. This is unsurprising because, prior to training, subjects have little-to-no competition for preverbal role assignments; after training, there is a preverbal extracted object category, which the model can erroneously use. This slight, though significant in Eve, deficit is counter-balanced by a very substantial and significant improvement in object-extraction labelling accuracy.

[41]Though performance is slightly worse when arguments are not collapsed, all the same patterns emerge.

Similarly, training confers a large and significant improvement for role assignment in wh-relative constructions, but it yields less of an improvement for that-relative constructions. This difference mimics a finding observed in the developmental literature where children seem slower to acquire comprehension of that-relatives than of wh-relatives (Gagliardi and Lidz, 2010).

6.6 Comparison to BabySRL

The acquisition of semantic role labelling (SRL) by the BabySRL model (Connor et al., 2008, 2009, 2010) bears many similarities to the current work and is, to our knowledge, the only comparable line of inquiry to the current one. The primary function of BabySRL is to model the acquisition of semantic role labelling while making an idiosyncratic error which infants also make (Gertner and Fisher, 2012), the 1-1 role bias error (John and Mary gorped interpreted as John gorped Mary). Similar to the model presented in this paper, BabySRL is based on simple ordering features such as argument position relative to the verb and argument position relative to the other arguments.

This section will demonstrate that the model in this paper initially reflects 1-1 role bias comparably to BabySRL, though it progresses beyond this bias after training.[42] Further, the model in this paper is able to reflect the concurrent acquisition of filler-gap whereas BabySRL does not seem well-suited to such a task. Finally, BabySRL performs undesirably in intransitive settings whereas the model in this paper does not.

[42] All evaluations in this section are preceded by training on the chunked Eve corpus.

Connor et al. (2008) demonstrate that a supervised perceptron classifier, based on positional features and trained on the silver role label annotations of the BabySRL corpus, manifests 1-1 role bias errors. Follow-up studies show that supervision may be lessened (Connor et al., 2009) or removed (Connor et al., 2010) and BabySRL will still reflect a substantial 1-1 role bias.

Connor et al. (2008) and Connor et al. (2009) run direct analyses of how frequently their models make 1-1 role bias errors. A comparable evaluation may be run on the current model by generating 1000 sentences with a structure of NNV and reporting how many times the model chooses a subject-first labelling (see Table 6.6).[43] The results of Connor et al. (2008) and Connor et al. (2009) depend on whether BabySRL uses argument-argument relative position as a feature or argument-verb relative position as a feature (there is no combined model). Further, the model presented here from Connor et al. (2009) has a unique argument constraint, similar to the model in this paper, in order to make comparison as direct as possible.

The 1-1 role bias error rate (before training) of the model presented in this paper is comparable to that of Connor et al. (2008) and Connor et al. (2009), which shows that the current model provides comparable developmental modeling benefits to the BabySRL models. Further, similar to real children (see Figure 6.1) the model presented in this paper develops beyond this error by the end of its training,[44] whereas the BabySRL models still make this error after training.

[43]While Table 6.6 analyzes erroneous labellings of NNV structure, the 'Obj' column of Table 6.5 (Left) shows model accuracy on NNV structures.

[44]It is important to note that the unique argument constraint prevents the current model from actually getting the correct, conjoined-subject parse, but it no longer exhibits agent-first bias, an important step for acquiring passives, which occurs between 3 and 4 years (Thatcher et al., 2008).

	Error rate
Initial	.36
Trained	.11
Initial (given 2 args)	.66
Trained (given 2 args)	.13
2008 arg-arg position	.65
2008 arg-verb position	0
2009 arg-arg position	.82
2009 arg-verb position	.63

Table 6.6: 1-1 role bias error in this model compared to the models of Connor et al. (2008) and Connor et al. (2009). That is, how frequently each model labelled an NNV sentence SOV. Since the Connor et al. models are perceptron-based, they require both arguments be labelled. The model presented in this paper does not share this restriction, so the raw error rate for this model is presented in the first two lines; the error rate once this additional restriction is imposed is given in the second two lines.

Connor et al. (2010) look at how frequently their model correctly labels the agent in transitive and intransitive sentences with unknown verbs (to demonstrate that it exhibits an agent-first bias). This evaluation can be replicated for the current study by generating 1,000 sentences with the transitive form of NVN and a further 1,000 sentences with the intransitive form of NV (see Table 6.7).

Since Connor et al. (2010) investigate the effects of different initial lexicons, this evaluation compares against the resulting BabySRL from each initializer: they initially seed their part-of-speech tagger with either the 10 or 365 most frequent nouns in the corpus or they dispense with the tagger and use gold part-of-speech tags.

As with subject extraction, the model in this paper gets less accurate after training because of the newly minted extracted object category that can be mistakenly used in these canonical settings. While the model of Connor et al. (2010) outperforms the

	NVN	NV
Sents in Eve	1173	1513
Sents in Adam	1029	1353
Initial	.67	1
Trained	.65	.96
Weak (10) lexical	.71	.59
Strong (365) lexical	.74	.41
Gold Args	.77	.58

Table 6.7: Agent-prediction recall accuracy in transitive (NVN) and intransitive (NV) settings of the model presented in this paper (middle) and the combined model of Connor et al. (2010) (bottom), which has features for argument-argument relative position as well as argument-predicate relative position and so is closest to the model presented in this paper.

model presented here when in a transitive setting, their model does much worse in an intransitive setting. The difference in transitive settings stems from increased lexicalization, as is apparent from their results alone; the model presented here initially performs close to their weakly lexicalized model, though training impedes agent-prediction accuracy due to an increased probability of non-canonical objects.

For the intransitive case, however, whereas the model presented in this paper is generally able to successfully label the lone noun as the subject, the model of Connor et al. (2010) chooses to label lone nouns as objects about 40% of the time. This likely stems from their model's reliance on argument-argument relative position as a feature; when there is no additional argument to use for reference, the model's accuracy decreases. This is borne out by their model (not shown in Table 6.7) that omits the argument-argument relative position feature and solely relies on verb-argument position, which achieves up to 70% accuracy in intransitive settings. Even in that case,

however, BabySRL still chooses to label lone nouns as objects 30% of the time. The fact that intransitive sentences are more common than transitive sentences in both the Eve and Adam sections of the BabySRL corpus suggests that learners should be more likely to assign correct roles in an intransitive setting, which is not reflected in the BabySRL results.

The overall reason for the different results between the current work and BabySRL is that BabySRL relies on positional features that measure the relative position of two individual elements (e.g. where a given noun is relative to the verb). Since the model in this paper operates over global orderings, it implicitly takes into account the positions of other nouns as it models argument position relative to the verb; object and subject are in competition as labels for preverbal nouns, so a preverbal object is usually only assigned once a subject has already been detected.

Further, while BabySRL consistently reflects 1-1 role bias (corresponding to a pre 25-month old learner), it also learns to productively label five roles, which developmental studies have shown does not take place until at least 31 months (Goldberg et al., 2004; Bello, 2012). Finally, it does not seem likely that BabySRL could be easily extended to capture filler-gap acquisition. The argument-verb position features impede acquisition of filler-gap by classifying preverbal arguments as agents, and the argument-argument position features inhibit accurate labelling in intransitive settings and result in an agent-first bias which would tend to label extracted objects as agents. In fact, these observations suggest that any linear classifier which relies on positioning features will have difficulties modeling filler-gap acquisition.

In sum, the unlexicalized model presented in this paper is able to achieve greater labelling accuracy than the lexicalized BabySRL models in intransitive settings, though

this model does perform slightly worse in the less common transitive setting. Further, the unsupervised model in this paper initially reflects developmental 1-1 role bias as well as the supervised BabySRL models, and it is able to progress beyond this bias. Finally, unlike BabySRL, the model presented here provides a cognitive model of the acquisition of filler-gap comprehension, which BabySRL does not seem well-suited to model.

6.7 Discussion

This paper has presented a simple cognitive model of filler-gap acquisition, which is able to capture several findings from developmental psychology. Training significantly improves role labelling in the case of object-extractions, which improves the overall accuracy of the model. This boost is accompanied by a slight decrease in labelling accuracy in subject-extraction settings. The asymmetric ease of subject versus object comprehension is well-documented in both children and adults (Gibson, 1998), and while training improves the model's ability to process object-extractions, there is still a gap between object-extraction and subject-extraction comprehension even after training.

Further, the model exhibits better comprehension of wh-relatives than that-relatives similar to children (Gagliardi and Lidz, 2010). This could also be an area where a lexicalized model could do better. As Gagliardi and Lidz (2010) point out, whereas wh-relatives such as who or which always signify a filler-gap construction, that can occur for many different reasons (demonstrative, determiner, complementizer, etc) and so is a much weaker filler-gap cue. A lexical model could potentially pick up

on clues which could indicate when that is a relativizer or simply improve on its comprehension of wh-relatives even more.

It is interesting to note that the current model does not make use of that as a cue at all and yet is still slower at acquiring that-relatives than wh-relatives. This fact suggests that the findings of Gagliardi and Lidz (2010) may be partially explained by a frequency effect: perhaps the input to children is simply biased such that wh-relatives are much more common than that-relatives (as shown in Table 6.5).

This model also initially reflects the 1-1 role bias observed in children (Gertner and Fisher, 2012) as well as previous models (Connor et al., 2008, 2009, 2010) without sacrificing accuracy in canonical intransitive settings.

Finally, this model is extremely robust to different initializations. The canonical Gaussian expectations can begin far from the verb (± 3) or close to the verb (± 0.1), and the standard deviations of the distributions and the skip-penalty can vary widely; the model always converges to give comparable results to those presented here. The only constraint on the initial parameters is that the probability of the extracted object occurring preverbally must exceed the skip-penalty (i.e. extraction must be possible). In short, this paper describes a simple, robust cognitive model of the development of a learner between 15 months until somewhere between 25- and 30-months old (since 1-1 role bias is no longer present but no more than two arguments are being generalized).

In future, it would be interesting to incorporate lexicalization into the model presented in this paper, as this feature seems likely to bridge the gap between this model and BabySRL in transitive settings. Lexicalization should also help further distinguish modifiers from arguments and improve the overall accuracy of the model.

It would also be interesting to investigate how well this model generalizes to languages besides English. Since the model is able to use the verb position as a semi-permeable boundary between canonical subjects and objects, it may not work as well in verb-final languages, and thus makes the prediction that filler-gap comprehension may be acquired later in development in such languages due to a greater reliance on hierarchical syntax.

Ordering is one of the definining characteristics of a language that must be acquired by learners (e.g. SVO vs SOV), and this work shows that filler-gap comprehension can be acquired as a by-product of learning orderings rather than having to resort to higher-order syntax. Note that this model cannot capture the constraints on filler-gap usage which require a hierarchical grammar (e.g. subjacency), but such knowledge is really only needed for successful production of filler-gap constructions, which occurs much later (around 5 years; de Villiers and Roeper, 1995). Further, the kind of ordering system proposed in this paper may form an initial basis for learning such grammars (Jackendoff and Wittenberg, in press).

Chapter 7: Conclusion

This work presented an accurate syntactic parser that estimates a variety of incremental complexity metrics based on syntactic occurrence frequencies (e.g., entropy reduction, PCFG surprisal). This work then identified an inconsistency in how surprisal is applied during reading time prediction. By summing n-gram surprisal over first-pass regions, the measure obtains a much better fit to reading times, even to the extent that syntactic PCFG surprisal becomes only weakly predictive of reading times.

However, this thesis demonstrated that predictive PCFG entropy correlates with reading times and that future PCFG surprisal may be used as a cheaper (though possibly incomplete) single-step predictive entropy approximation than the hallucinatory point-wise approximation that is more commonly used (e.g., by the Roark, 2001, parser). Further, this work has shown that approximating entropy with a finer-grained language model provides a better correlation to reading times than coarser language models, necessitating the use of surprisal as an entropy approximation. In addition, this work corroborated findings by Angele et al. (2015) that uncertainty about lexical frequencies only seems to extend to the one or two words following a fixated word, suggesting that single-step predictive entropy has theoretical validity

as a reading time predictor, and highlighting a possible confound for studies looking for parafoveal-on-foveal processing effects.

Finally, this work presented a proof-of-concept statistical long-distance dependency acquisition model that can unify findings from the child language acquisition literature by replicating the developmental timeline of young English learners. Such a model could provide the basis for a system that could ultimately be used to acquire hierarchical syntax through an initial sensitivity to linear lexical frequencies.

In sum, this thesis has presented evidence that lexical and syntactic frequencies independently influence predictions of upcoming material and independently influence reading times of observed material. This work has provided a tool for easily estimating those influences and has suggested a means by which hierarchical syntax could be acquired by humans through sensitivity to linear lexical frequencies.

Bibliography

Steven P. Abney and Mark Johnson. Memory requirements and local ambiguities of parsing strategies. J. Psycholinguistic Research, 20(3):233–250, 1991.

Nameera Akhtar. Acquiring basic word order: evidence for data-driven learning of syntactic structure. Journal of Child Language, 26:339–356, 1999.

Gerry T. M. Altmann and Yuki Kamide. Incremental interpretation at verbs: Restricting the domain of subsequent reference. Cognition, 73(3):247–264, 1999.

Bernhard Angele, Elizabeth R. Schotter, Timothy J. Slattery, Tara L. Tenenbaum, Klinton Bicknell, and Keith Rayner. Do successor effects in reading reflect lexical parafoveal processing? evidence from corpus-based and experimental eye movement data. Journal of Memory and Language, 79–80:76–96, 2015.

Manabu Arai and Frank Keller. The use of verb-specific information for prediction in sentence processing. Language and Cognitive Processes, 28(4):525–560, 2013.

Fred Attneave. Applications of Information Theory to Psychology: A summary of basic concepts, methods and results. Holt, Rinehart, and Winston, 1959.

Matthew Aylett and Alice Turk. Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. Journal of the acoustical society of America, 119(5):3048–3059, 2006.

Emmon Bach. Discontinuous constituents in generalized categorial grammars. Proceedings of the Annual Meeting of the Northeast Linguistic Society (NELS), 11: 1–12, 1981.

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. lme4: Linear mixed-effects models using Eigen and S4, 2014. URL http://CRAN.R-project.org/package=lme4. R package version 1.1-7.

Richard Bellman. Dynamic Programming. Princeton University Press, Princeton, NJ, 1957.

Sophia Bello. Identifying indirect objects in French: An elicitation task. In Proceedings of the 2012 annual conference of the Canadian Linguistic Association, 2012.

Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly, Beijing, 2009. ISBN 978-0-596-51649-9.

R. A. Bjork and W. B. Whitten. Recency sensitive retrieval processes in long-term free recall. Cognitive Psychology, 6:173–189, 1974.

Paul Boersma. How we learn variation, optionality, and probability. Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam, 21:43–58, 1997.

Taylor L. Booth and Richard A. Thompson. Applying probability measures to abstract languages. IEEE Transactions on Computers, C-22(5):442–450, 1973.

Matthew Botvinick. Multilevel structure in behavior and in the brain: a computational model of Fuster's hierarchy. Philosophical Transactions of the Royal Society, Series B: Biological Sciences, 362:1615–1626, 2007.

G. E. P. Box and D. R. Cox. An analysis of transformations. Journal of the Royal Statistical Society, B, 26:211–234, 1964.

Harr Chen, S.R.K. Branavan, Regina Barzilay, and David R. Karger. Content modeling using latent permutations. Journal of Artificial Intelligence Research, 36: 129–163, 2009.

Noam Chomsky and George A. Miller. Introduction to the formal analysis of natural languages. In Handbook of Mathematical Psychology, pages 269–321. Wiley, New York, NY, 1963.

Michael Collins. Three generative, lexicalised models for statistical parsing. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97), 1997.

Michael Connor, Yael Gertner, Cynthia Fisher, and Dan Roth. Baby srl: Modeling early language acquisition. In Proceedings of the Twelfth Conference on Computational Natural Language Learning, 2008.

Michael Connor, Yael Gertner, Cynthia Fisher, and Dan Roth. Minimally supervised model of early language acquisition. In Proceedings of the Thirteenth Conference on Computational Natural Language Learning, 2009.

Michael Connor, Yael Gertner, Cynthia Fisher, and Dan Roth. Starting from scratch in semantic role labelling. In Proceedings of ACL 2010, 2010.

Matthew Crocker and Thorsten Brants. Wide-coverage probabilistic sentence processing. Journal of Psycholinguistic Research, 29(6):647–669, 2000.

Robert G. Crowder. The demise of short-term memory. Acta Psychologica, 50(3): 291–323, 1982.

Peter Culicover. Explaining syntax: representations, structures, and computation. Oxford University Press, 2013.

Jill de Villiers and Thomas Roeper. Barriers, binding, and acquisition of the dp-np distinction. Language Acquisition, 4(1):73–104, 1995.

Katherine A DeLong, Melissa Troyer, and Marta Kutas. Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. Language and Linguistics Compass, 8(12):631–645, 2014.

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

Vera Demberg, Asad B. Sayeed, Philip J. Gorinski, and Nikolaos Engonopoulos. Syntactic surprisal affects spoken word duration in conversational contexts. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 356–367, 2012.

Vera Demberg, Frank Keller, and Alexander Koller. Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. Computational Linguistics, 39(4):1025–1066, 2013.

Holger Diessel and Michael Tomasello. The acquisition of finite complement clauses in English: A corpus-based analysis. Cognitive Linguistics, 12:1–45, 2001.

Alex B Fine, T Florian Jaeger, Thomas A Farmer, and Ting Qian. Rapid expectation adaptation during syntactic comprehension. PloS ONE, 8(10):1–18, 2013.

Victoria Fossum and Roger Levy. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In Proceedings of CMCL 2012. Association for Computational Linguistics, 2012.

Stefan Frank and Rens Bod. Insensitivity of the human sentence-processing system to hierarchical structure. Psychological Science, 22:829–834, 2011.

Stefan L. Frank. Word embedding distance does not predict word reading time. In Proc. Annual Meeting of the Cognitive Science Society, 2017.

Stefan L. Frank, Rens Bod, and Morton H. Christiansen. How hierarchical is language use? Proceedings of the Royal Society B, 279:4522–4531, 2012.

Stefan L. Frank, Irene Fernandez Monsalve, Robin L. Thompson, and Gabriella Vigliocco. Reading time data for evaluating broad-coverage models of English sentence processing. Behavior Research Methods, 45:1182–1190, 2013.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. The ERP response to the amount of information conveyed by words in sentences. Brain & Language, 140:1–11, 2015.

Stefan L. Frank, Thijs Trompenaars, and Shravan Vasishth. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? Cognitive Science, 40:554–578, 2016.

Lyn Frazier. Sentence processing: A tutorial review. In M. Coltheart, editor, Attention and Performance 12: The Psychology of Reading, pages 559–586. Erlbaum, Hillsdale, NJ, 1987.

Lyn Frazier and Charles Clifton, Jr. Construal. MIT Press, Cambridge, MA, 1996.

Lyn Frazier and Keith Rayner. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. Cognitive Psychology, 14:178–210, 1982.

Richard Futrell, Edward Gibson, Hal Tily, Anastasia Vishnevetsky, Steve Piantadosi, and Evelina Fedorenko. Natural stories corpus. in prep.

Annie Gagliardi and Jeffrey Lidz. Morphosyntactic cues impact filler-gap dependency resolution in 20- and 30-month-olds. In Poster session of BUCLD35, 2010.

Annie Gagliardi, Tara M. Mease, and Jeffrey Lidz. Discontinuous development in the acquisition of filler-gap dependencies: Evidence from 15- and 20-month-olds. Harvard unpublished manuscript: http://www.people.fas.harvard.edu/∼gagliardi, in prep.

Gerald Gazdar, Ewan Klein, Geoffrey Pullum, and Ivan Sag. Generalized Phrase Structure Grammar. Harvard University Press, Cambridge, MA, 1985.

Yael Gertner and Cynthia Fisher. Predicted errors in children's early sentence comprehension. Cognition, 124:85–94, 2012.

Edward Gibson. A computational theory of human linguistic processing: Memory limitations and processing breakdown. PhD thesis, Carnegie Mellon, 1991.

Edward Gibson. Linguistic complexity: Locality of syntactic dependencies. Cognition, 68(1):1–76, 1998.

Edward Gibson. The dependency locality theory: A distance-based theory of linguistic complexity. In Image, language, brain: Papers from the first mind articulation project symposium, pages 95–126, Cambridge, MA, 2000. MIT Press.

Lila R. Gleitman. The structural sources of verb meanings. Language Acquisition, 1: 3–55, 1990.

Adele E. Goldberg, Devin Casenhiser, and Nitya Sethuraman. Learning argument structure generalizations. Cognitive Linguistics, 14(3):289–316, 2004.

Sharon Goldwater and Tom Griffiths. A fully Bayesian approach to unsupervised part-of-speech tagging. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, 2007.

David Graff and Christopher Cieri. English Gigaword LDC2003T05, 2003.

John Hale. A probabilistic earley parser as a psycholinguistic model. In Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics, pages 159–166, Pittsburgh, PA, 2001.

John Hale. Grammar, Uncertainty and Sentence Processing. PhD thesis, Cognitive Science, The Johns Hopkins University, 2003.

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

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 690–696, Sofia, Bulgaria, August 2013.

James Henderson. Lookahead in deterministic left-corner parsing. In Proc. Workshop on Incremental Parsing: Bringing Engineering and Cognition Together, pages 26–33, Barcelona, Spain, 2004.

Marc W. Howard and Michael J. Kahana. A distributed representation of temporal context. Journal of Mathematical Psychology, 45:269–299, 2002.

Kiwako Ito and Shari R. Speer. Anticipatory effect of intonation: Eye movements during instructed visual search. Journal of Memory and Language, 58:541–573, 2008.

Ray Jackendoff and Eva Wittenberg. What you can say without syntax: A hierarchy of grammatical complexity. In Fritz Newmeyer and Lauren Preston, editors, Measuring Linguistic Complexity. Oxford University Press, in press.

Philip N. Johnson-Laird. Mental models: Towards a cognitive science of language, inference, and consciousness. Harvard University Press, Cambridge, MA, USA, 1983. ISBN 0-674-56882-6.

Aravind K. Joshi, K. Vijay Shanker, and David Weir. The convergence of mildly context-sensitive grammar formalisms. Technical Report MS-CIS-90-01, Department of Computer and Information Science, University of Pennsylvania, January 1990.

Daniel Jurafsky. A probabilistic model of lexical and syntactic access and disambiguation. Cognitive Science: A Multidisciplinary Journal, 20(2):137–194, 1996.

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D. Raymond. Probabilistic relations between words: Evidence from reduction in lexical production. In

Joan Bybee and Paul Hopper, editors, Frequency and the emergence of linguistic structure, pages 229–254. John Benjamins, Amsterdam, 2001.

Yuki Kamide, Gerry T. M. Altmann, and Sarah L Haywood. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and Language, 49(1):133–156, 2003a.

Yuki Kamide, Christoph Scheepers, and Gerry T. M. Altmann. Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. Journal of Psycholinguistic Research, 32(1):37–55, 2003b.

Alan Kennedy, James Pynte, and Robin Hill. The Dundee corpus. In Proceedings of the 12th European conference on eye movement, 2003.

Albert Kim and Vicky Lai. Rapid interactions between lexical semantic and word form analysis during word recognition in context: Evidence from erps. Journal of Cognitive Neuroscience, 24(5):1104–1112, 2012.

Dan Klein and Christopher D. Manning. Corpus-based induction of syntactic structure: Models of dependency and constituency. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 2004.

M Kutas and S A Hillyard. Reading senseless sentences: brain potentials reflect semantic incongruity. Science, 207(4427):203–205, jan 1980. doi: 10.1126/science.7350657.

Tom Kwiatkowski, Sharon Goldwater, Luke S. Zettlemoyer, and Mark Steedman. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In Proceedings of EACL 2012, 2012.

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

Richard L. Lewis and Shravan Vasishth. An activation-based model of sentence processing as skilled memory retrieval. Cognitive Science, 29(3):375–419, 2005.

Tal Linzen and T. Florian Jaeger. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. Cognitive Science, pages 1–30, 2015.

Steven G Luke and Kiel Christiansen. Predicting inflectional morphology from context. Language, Cognition and Neuroscience, pages 1–14, 2015.

Brian MacWhinney. The CHILDES project: Tools for analyzing talk. Lawrence Elrbaum Associates, Mahwah, NJ, third edition, 2000.

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

David Marr. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W.H. Freeman and Company, 1982.

David McClosky. Any Domain Parsing: Automatic Domain Adapatation for Parsing. PhD thesis, Computer Science Department, Brown University, 2010.

George Miller and Noam Chomsky. Finitary models of language users. In R. Luce, R. Bush, and E. Galanter, editors, Handbook of Mathematical Psychology, volume 2, pages 419–491. John Wiley, 1963.

Don C. Mitchell. Lexical guidance in human parsing: Locus and processing characteristics. In M. Coltheart, editor, Attention and performance XII: The Psychology of Reading, pages 601–618. Erlbaum, Hillsdale, NJ, 1987.

Letitia R. Naigles. Children use syntax to learn verb meanings. The Journal of Child Language, 17:357–374, 1990.

Luan Nguyen, Marten van Schijndel, and William Schuler. Accurate unbounded dependency recovery using generalized categorial grammars. In Proceedings of the 24th International Conference on Computational Linguistics (COLING '12), pages 2125–2140, Mumbai, India, 2012.

Slav Petrov and Dan Klein. Improved inference for unlexicalized parsing. In Proceedings of NAACL HLT 2007, pages 404–411, Rochester, New York, April 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N/N07/N07-1051.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL'06), 2006.

Colin Phillips. Some arguments and non-arguments for reductionist accounts of syntactic phenomena. Language and Cognitive Processes, 28:156–187, 2010.

Martin Pickering and Guy Barry. Sentence processing without empty categories. Language and Cognitive Processes, 6(3):229–259, 1991.

Martin J. Pickering and Matthew J. Traxler. Evidence against the use of subcategorisation frequency in the processing of unbounded dependencies. Language and Cognitive Processes, 18(4):469–503, 2003.

Martin J. Pickering, Matthew J. Traxler, and Matthew W. Crocker. Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. Journal of Memory and Language, 43:447–475, 2000.

Carl Pollard and Ivan Sag. Head-driven Phrase Structure Grammar. University of Chicago Press, Chicago, 1994.

Vasin Punyakanok, Dan Roth, and Wen-tau Yih. The importance of syntactic parsing and inference in semantic role labeling. Computational Linguistics, 34(2):257–287, 2008. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/coli.2008.34.2.257.

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

Brian Roark. Expected surprisal and entropy. Technical Report CSLU-11-004, Center for Spoken Language Processing, Oregon Health and Science University, Portland, OR, 2011.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. Proceedings of the 2009 Conference on Empirical Methods in Natural Langauge Processing, pages 324–333, 2009.

S. J. Rosencrantz and P. M. Lewis II. Deterministic left corner parsing. In Proceedings of the IEEE Conference Record of the 11th Annual Symposium on Switching and Automata, pages 139–152, 1970.

John R. Ross. Constraints on Variables in Syntax. PhD thesis, Massachusetts Institute of Technology, 1967.

William Schuler. Effects of filler-gap dependencies on working memory requirements for parsing. In Proceedings of COGSCI, pages 501–506, Austin, TX, 2011. Cognitive Science Society.

William Schuler, Samir AbdelRahman, Tim Miller, and Lane Schwartz. Broad-coverage incremental parsing using human-like memory constraints. Computational Linguistics, 36(1):1–30, 2010.

Amanda Seidl, George Hollich, and Peter W. Jusczyk. Early understanding of subject and object wh-questions. Infancy, 4(3):423–436, 2003.

Satoshi Sekine. The domain dependence of parsing. In Proceedings of the Fifth Conference on Applied Natural Language Processing, pages 96–102. Association for Computational Linguistics, 1997.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. Memory access during incremental sentence processing causes reading time latency. In Proceedings of the Computational Linguistics for Linguistic Complexity Workshop. Association for Computational Linguistics, 2016.

Claude Shannon. A mathematical theory of communication. Bell System Technical Journal, 27:379–423, 623–656, 1948.

Rushen Shi, Janet F. Werker, and James L. Morgan. Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. Cognition, 72(2):B11–B21, 1999.

Timothy J. Slattery, Patrick Sturt, Kiel Christianson, Masaya Yoshida, and Fernanda Ferreira. Lingering misinterpretations of garden path sentences arise from competing syntactic representations. Journal of Memory and Language, 69:104–120, 2013.

Nathaniel J. Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. Cognition, 128:302–319, 2013.

Adrian Staub. The parser doesn't ignore intransitivity, after all. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(3):550–569, 2007.

Adrian Staub, Charles Clifton, and Lyn Frazier. Heavy NP shift is the parser's last resort: Evidence from eye movements. Journal of Memory and Language, 54: 389–406, 2006.

Mark Steedman. The syntactic process. MIT Press/Bradford Books, Cambridge, MA, 2000.

Patrick Sturt and Vincent Lombardo. Processing coordinate structures: Incrementality and connectedness. Cognitive Science, 29:291–305, 2005.

Katherine Thatcher, Holly Branigan, Janet McLean, and Antonella Sorace. Children's early acquisition of the passive: Evidence from syntactic priming. In Proceedings of the Child Language Seminar 2007, pages 195–205, University of Reading, 2008.

Ivan Titov and Alexandre Klementiev. Crosslingual induction of semantic roles. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL-2011), 2012.

Jos J A Van Berkum, Colin M Brown, Pienie Zwitserlood, Valesca Kooijman, and Peter Hagoort. Anticipating upcoming words in discourse: evidence from erps and reading times. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(3):443, 2005.

R. P. G. van Gompel and Martin J. Pickering. Lexical guidance in sentence processing: A note on Adams, Clifton, and Mitchell (1998). Psychonomic Bulletin and Review, 8:851–857, 2001.

Marten van Schijndel and Micha Elsner. Bootstrapping into filler-gap: An acquisition story. In Fifty-Second Annual Meeting of the Association for Computational Linguistics (ACL 2014), 2014.

Marten van Schijndel and William Schuler. An analysis of frequency- and memory-based processing costs. In Proceedings of NAACL-HLT 2013. Association for Computational Linguistics, 2013.

Marten van Schijndel and William Schuler. Hierarchic syntax improves reading time prediction. In Proceedings of NAACL-HLT 2015. Association for Computational Linguistics, 2015.

Marten van Schijndel and William Schuler. Addressing surprisal deficiencies in reading time models. In Proceedings of the Computational Linguistics for Linguistic Complexity Workshop. Association for Computational Linguistics, 2016.

Marten van Schijndel and William Schuler. Approximations of predictive entropy correlate with reading times. In Proc. of CogSci 2017. Cognitive Science Society, 2017.

Marten van Schijndel, Andy Exley, and William Schuler. A model of language processing as hierarchic sequential prediction. Topics in Cognitive Science, 5(3):522–540, 2013a.

Marten van Schijndel, Luan Nguyen, and William Schuler. An analysis of memory-based processing costs using incremental deep syntactic dependency parsing. In Proc. of CMCL 2013. Association for Computational Linguistics, 2013b.

Marten van Schijndel, William Schuler, and Peter W Culicover. Frequency effects in the processing of unbounded dependencies. In Proc. of CogSci 2014. Cognitive Science Society, 2014.

Sandra R. Waxman and Amy E. Booth. Seeing pink elephants: Fourteen-month-olds' interpretations of novel nouns and adjectives. Cognitive Psychology, 43:217–242, 2001.

Nicole Y Y Wicha, Eva M Moreno, and Marta Kutas. Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in spanish sentence reading. Journal of Cognitive Neuroscience, 16(7):1272–1288, 2004.

K. Wilson and J. B. Carroll. Applications of entropy measures to problems of sequential structure. In C. E. Osgood and T. A. Sebeok, editors, Psycholinguistics:

A survey of theory and research, pages 103–110. Indiana University Press, Bloomington, 1954.

Stephen Wu, Asaf Bachrach, Carlos Cardenas, and William Schuler. Complexity metrics in an incremental right-corner parser. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10), pages 1189–1198, 2010.

Victor H. Yngve. A Model and an Hypothesis for Language Structure. Proceedings of the American Philosophical Society, 104(5):444–466, 1960.

Sylvia Yuan, Cynthia Fisher, and Jesse Snedeker. Counting the nouns: Simple structural cues to verb meaning. Child Development, 83(4):1382–1399, 2012.