A Large-Scale Study of the Effects of Word Frequency and Predictability in Naturalistic Reading

Cory Shain
Department of Linguistics
The Ohio State University
shain.3@osu.edu

Abstract

A number of psycholinguistic studies have factorially manipulated words' contextual predictabilities and corpus frequencies and shown separable effects of each on measures of human sentence processing, a pattern which has been used to support distinct mechanisms underlying prediction on the one hand and lexical retrieval on the other. This paper examines the generalizability of this finding to more realistic conditions of sentence processing by studying effects of frequency and predictability in three large-scale naturalistic reading corpora. Results show significant effects of word frequency and predictability in isolation but no effect of frequency over and above predictability, and thus do not provide evidence of distinct mechanisms. The non-replication of separable effects in a naturalistic setting raises doubts about the existence of such a distinction in everyday sentence comprehension. Instead, these results are consistent with previous claims that apparent effects of frequency are underlyingly effects of predictability.

1 Introduction

Are there distinct effects of a word's frequency versus predictability in human sentence comprehension? Recent evidence implicates prediction as a major organizing principle in cognition (Bubic et al., 2010; Singer et al., 2018; Keller and Mrsic-Flogel, 2018), and psycholinguists have long studied the role of prediction in human sentence processing and its relation to other comprehension mechanisms (Marslen-Wilson, 1975; Kutas and Hillyard, 1984; MacDonald et al., 1994; Tanenhaus et al., 1995; Hale, 2001; Norris, 2006; Levy, 2008; Frank and Bod, 2011). Some prominent theories of word recognition claim that ease of lexical access is modulated by the strength of a word's representation in memory, independently of contextual factors that guide prediction (Seidenberg and McClelland, 1989; Coltheart et al., 2001; Harm and Seidenberg, 2004). Other theories hold that apparent effects of frequency are underlyingly effects of predictability (Norris, 2006; Levy, 2008; Rasmussen and Schuler, 2018).

A number of studies using constructed stimuli that factorially manipulate word frequency and predictability have found separable additive effects of each, suggesting distinct influences on lexical processing (see Staub, 2015 for a review). This paper examines the generalizability of these findings to typical sentence comprehension by searching for separable effects of frequency and n-gram predictability using deconvolutional time series regression (DTSR) models (Shain and Schuler, 2018) fitted to three large naturalistic reading corpora: Natural Stories (Futrell et al., 2018), Dundee (Kennedy et al., 2003), and UCL (Frank et al., 2013). While results show evidence of both frequency and predictability effects in isolation, they show no effect of frequency over predictability and thus do not support the existence of separable effects. They are instead consistent with either (1) an account of apparent frequency effects as epiphenomena of predictive processing (Norris, 2006; Levy, 2008) or (2) a more circumscribed role for frequency effects in naturalistic reading than constructed experiments suggest.

2 Background and Related Work

2.1 Frequency and Predictability in Human Sentence Processing

It has long been recognized that low-frequency words are harder to process (Inhoff and Rayner, 1986). For example, in a neutral context, the more frequent *bottle* should on average be processed more quickly than the less frequent *kettle*:

Proceedings of NAACL-HLT 2019, pages 4086–4094
Minneapolis, Minnesota, June 2 - June 7, 2019.

(1) a. I have a **bottle**.
b. I have a **kettle**.

However, context can dramatically alter these patterns by changing words' predictability (Ehrlich and Rayner, 1981):

(2) a. the pot calling the **bottle** black
b. the pot calling the **kettle** black

Some models of word recognition (Seidenberg and McClelland, 1989; Coltheart et al., 2001; Harm and Seidenberg, 2004) posit a context-independent lexical retrieval mechanism, distinct from any mechanisms for predictive coding, with processing cost proportional to the strength of a word's representation in memory (a function of lexical frequency). Such a view predicts separable effects of frequency and predictability in human language comprehension. Other models (Hale, 2001; Norris, 2006; Levy, 2008; Rasmussen and Schuler, 2018) posit no such context-independent retrieval mechanism, and instead propose a unified comprehension mechanism that incrementally reallocates resources between possible interpretations of the unfolding sentence, with processing cost proportional to the amount of information (resource reallocation) contributed by each new word. Such a view predicts no separable effects of frequency and predictability because lexical frequencies are subsumed into the incremental probability model.

Consistently with the first hypothesis, previous studies have shown separable additive effects of frequency and predictability by factorially manipulating corpus frequency and cloze predictability (Rayner et al., 2004; Ashby et al., 2005; Gollan et al., 2011; Staub and Benatar, 2013, see Staub, 2015 for a review). However, cloze estimates poorly distinguish degrees of low contextual probability (Smith and Levy, 2013), and constructed stimuli, while affording direct control over linguistic variables, may fail to reflect the typical distributional characteristics of the language, lack context, and/or inadvertently trigger suspension of the usual processes of pragmatic inference due to the absence of an overarching discourse (Demberg and Keller, 2008; Hasson and Honey, 2012; Shain et al., 2018). It is therefore not yet clear whether frequency and predictability effects can be separated in a more realistic setting.

2.2 The Naturalistic Experimental Paradigm

Concerns about the ecological validity of constructed stimuli can be addressed by the use of naturalistic stimuli (e.g. stories, newspaper articles, persuasive pieces, etc.). Naturalistic experiments are therefore an important complement to constructed experiments in the study of cognitive processes (Hasson and Honey, 2012).

However, naturalistic experiments introduce their own challenges. Without the ability to factorially manipulate frequency and predictability, naturalistic studies must confront the natural collinearity between these two variables in ordinary language (Demberg and Keller, 2008). Furthermore, because naturalistic stimuli do not define a critical region of the stimulus, responses are generally modeled word-by-word (Demberg and Keller, 2008; Frank and Bod, 2011; Smith and Levy, 2013; van Schijndel and Schuler, 2015). It is standard psycholinguistic practice to do so through ablative likelihood ratio testing (LRT) of linear mixed effects regression (LMER) models (Bates et al., 2015) fitted to the dependent variable of interest (e.g. fixation duration) (Demberg and Keller, 2008; Frank and Bod, 2011; van Schijndel and Schuler, 2015; Shain et al., 2016). However, this approach has important disadvantages. First, naturalistic data constitute time series that may violate the independence assumptions of linear regression and therefore confound model interpretation and hypothesis testing (Baayen et al., 2017, 2018; Shain and Schuler, 2018). One major such confound is *temporal diffusion* (i.e. a lingering response to stimuli), which can be brought under statistical control through deconvolutional time series regression (DTSR) models that directly estimate temporal structure in the relationships between predictors and response (Shain and Schuler, 2018). Second, LRT implicitly evaluates on in-sample data, making it challenging to diagnose overfitting and to assess external validity (Vasishth et al., 2018). This can be addressed through out-of-sample non-parametric tests, such as the paired permutation test widely used in machine learning (Demšar, 2006).

3 Experimental Setup

This paper seeks to complement constructed stimulus experiments by searching for separable effects of frequency and predictability during naturalistic reading, using methods designed to ad-

Corpus	Effect estimate (log-ms)							
	SentPos	Trial	Rate	WordLen	SaccLen	PrevFix	Unigram	5-gram
Natural Stories	0.0098	-0.0216	-0.3069	—	—	0.0158	-0.0018	0.0174
Dundee	-0.0085	-0.0052	-0.0277	0.0068	-0.0021	-0.0178	-0.0067	0.0117
UCL	0.0524		-0.1330	0.0023	0.0221	0.0778	0.0005	0.0184

Table 1: Effect estimates in log-ms by corpus, computed as the IRF integral over the longest time offset seen in training. Following psycholinguistic convention, unigrams and 5-grams have opposite sign (log prob vs. surprisal). In UCL, *sentence position* and *trial* are identical (sentences were shuffled).

Corpus	ρ
Natural Stories	-0.78
Dundee	-0.73
UCL	-0.74

Table 2: Pearson's correlation between *5-gram surprisal* and *unigram log probability* by corpus.

dress the challenges of Section 2.2. The problem of temporal diffusion is addressed by using DTSR models rather than LMER (see Appendix A for implementation details). The problem of external validity is addressed by using held-out paired permutation testing rather than LRT, thus basing the hypothesis test directly on generalization error. The possibility that cloze probabilities are poor estimates of predictability for low-frequency words is addressed by operationalizing predictability as 5-gram surprisal generated by a large-vocabulary statistical language model. The natural collinearity of frequency and predictability is addressed through the use of large-scale data that should permit subtle differentiation of collinear effects. Taken together, the corpora examined in this study contain over one million fixations generated by 243 human subjects. Although there is a large-magnitude correlation between *unigram log probability* (frequency) and *5-gram surprisal* (predictability) in these corpora, as shown in Table 2, synthetic experiments show that DTSR can faithfully identify models from much smaller data than that used here, even when all predictors are correlated at the 0.75 level (Shain, 2018). Given the size of the data, failure to distinguish effects of frequency and predictability would raise doubts about the existence of such a separation in naturalistic reading.

3.1 Statistical Procedure

DTSR models are fitted separately to each of the Natural Stories (Futrell et al., 2018), Dundee (Kennedy et al., 2003), and UCL (Frank et al., 2013) corpora.[1] Following previous investigations of this question (Rayner et al., 2004; Ashby et al., 2005; Gollan et al., 2011, *inter alia*), frequency is estimated from corpus statistics — in this case, KenLM (Heafield et al., 2013) unigram models trained on the Gigaword 3 corpus (Graff and Cieri, 2003). Unlike previous studies using close estimates of predictability (Rayner et al., 2004; Ashby et al., 2005; Gollan et al., 2011, *inter alia*), predictability is statistically estimated, again using KenLM models (5-gram) trained on Gigaword 3. This is both because (1) cloze norming all words contained in thousands of naturalistic sentences is prohibitive and (2) statistical language models trained on large data can more reliably differentiate low probability continuations (Smith and Levy, 2013). Following recent work on prediction effects in naturalistic sentence comprehension (Demberg and Keller, 2008; Frank and Bod, 2011; Smith and Levy, 2013), predictability estimates are encoded as *surprisal* by negating the 5-gram log probabilities.

The models assume ShiftedGamma impulse response functions (Shain and Schuler, 2018, see Appendix A) for each of these variables, as well as for the nuisance variables *word length*, *saccade length* and an indicator variable for whether the previous word was fixated.[2] To capture trends in the response at different timescales, the mod-

[1] Natural Stories is a self-paced reading corpus containing 848,768 word fixations from 181 subjects reading narrative and informational texts. Dundee is an eye-tracking corpus containing 260,065 word fixations from 10 subjects reading newspaper editorials. UCL is an eye-tracking corpus containing 53,070 fixations from 42 subjects reading sentences taken from novels by amateur authors.

Although the sentences in UCL were randomized and presented in isolation — and therefore subject to some of the concerns about constructed stimuli raised in Section 2 — they are included here because the stimuli are naturally occurring rather than constructed for a particular experimental purpose. The UCL results replicate the overall pattern of significance (Table 3), and excluding them has no impact on the overall results.

[2] The variables *saccade length* and *previous was fixated* are only used for eye-tracking since they are not relevant to self-paced reading.

Comparison	Pooled	Corpus		
		Natural Stories	Dundee	UCL
5-gram only vs. baseline	**0.0001*** **	**0.0001*** **	**0.0001*** **	**0.0001*** **
Unigram only vs. baseline	**0.0001*** **	**0.0001*** **	**0.0001*** **	**0.0001*** **
5-gram + Unigram vs. Unigram-only	**0.0001*** **	**0.0001*** **	0.0626	**0.0006*** **
5-gram + Unigram vs. 5-gram-only	0.1515	0.1831	0.0105	0.1491

Table 3: Held-out paired permutation testing results, both pooled (left) and by corpus (right).

els also include linear effects for the word's index in the sentence (*sentence position*) and document (*trial*). Following Shain and Schuler (2018), in addition to the intercept, the models contain a convolved intercept (*rate*) designed to capture effects of stimulus timing. The response used in all corpora is log fixation duration (go-past for eye-tracking).[3] Outlier filtering is performed in each corpus following the procedures described in Shain and Schuler (2018).

Approximately half the data in each corpus is used for training, with the remaining half reserved for held-out evaluation. Models include by-subject random intercepts as well as by-subject random slopes and impulse response parameters for each predictor.[4] Held-out hypothesis testing uses a "diamond" ablative structure first ablating fixed effects for *5-gram surprisal* and *unigram log probability* individually and then ablating both. All random effects are retained in all models. Comparisons use paired permutation tests of the by-item losses on the evaluation set, pooling across all corpora.[5] Note that the non-parametric permutation test permits this pooling procedure to unify the models from all three corpora into a single test, since (unlike LRT) permutation testing supports out-of-sample comparison. Data processing was performed using the ModelBlocks toolchain (van Schijndel and Schuler, 2013), available at `https://github.com/modelblocks/modelblocks-release`. Model fitting was performed using the DTSR software library (Shain and Schuler, 2018), available at `https://github.com/coryshain/dtsr`. See the citations above for data access instructions.

[3]The overall pattern of significance does not change when first-pass durations are used.

[4]By-word random intercepts are not included because of their potential to subsume frequency effects.

[5]To correct for different error variances, errors are rescaled by the joint standard deviation of the errors from the full and ablated models by corpus.

4 Results

Effect estimates[6] from the full models are presented in Table 1 and pooled statistical comparisons are presented in the *Pooled* column of Table 3. If predictability and frequency effects are additive, all four comparisons in Table 3 should be significant. As shown, this is not the case. There is evidence that both frequency (*unigram log probability*) and predictability (*5-gram surprisal*) in isolation reliably index processing difficulty, as shown by the significance of both effects over the baseline. However, when the effects are compared to each other, predictability explains significantly more variance than frequency but not vice versa.

This general pattern of results further obtains for each corpus individually, as shown by the *Corpus* column breakdown in Table 3. One minor exception is that neither predictability nor frequency improves significantly over the other in Dundee.[7] The Dundee results are nevertheless consistent with an interpretation in which frequency and predictability do not index distinct processing phenomena and inconsistent with an interpretation in which they do. These results thus provide no evidence of separable frequency and predictability effects, whether the corpora are considered together or individually.

5 Discussion

As described in Section 4, results show no evidence of separable effects of frequency and predictability in naturalistic reading. One possible explanation for this outcome is that 5-gram surprisal tracks human prediction effort better than cloze probabilities, in part because cloze probabilities are less reliable for infrequent words. Although countervailing evidence exists in the literature (e.g. Smith and Levy, 2011 found effects of cloze but not n-gram probabilities in human read-

[6] The estimated impulse response functions that underlie these effect sizes are plotted in Appendix B.

[7]The p-value of 0.0105 observed for frequency over predictability does not achieve significance at the 0.05 level under 6-way Bonferroni correction (2 variables $\times$ 3 corpora).

ing times), in general this evidence is based on weak statistical competitors to cloze (e.g. Smith and Levy, 2011 used tri-grams). By contrast, recent trends in cognitive modeling point toward a correlation between the linguistic and psycholinguistic performance of language models, such that more powerful models with lower perplexity also tend to correlate more strongly with measures of cognitive effort (Goodkind and Bicknell, 2018; van Schijndel and Linzen, 2018). This suggests that apparent frequency effects may arise in part from poor estimates of predictability. Note that by using 5-gram surprisal rather than more powerful neural language models (Jozefowicz et al., 2016), the analysis described in this paper is conservative in its attribution of variance to predictability. The failure of frequency is thus all the more compelling, since replacing 5-gram surprisal with surprisals obtained from more powerful language models would be unlikely to increase the explanatory power of frequency.

Another potential explanation for the lack of separable effects of frequency and predictability is the use of naturalistic rather than constructed stimuli. Neuroscientific evidence shows that domain-general executive control regions activate during the processing of some artificially constructed language stimuli (Kaan and Swaab, 2002; Kuperberg et al., 2003; Novick et al., 2005; January et al., 2009) but fail to activate during the processing of naturalistic stimuli (Blank and Fedorenko, 2017). Such results have led some to argue that artificially constructed experimental stimuli may increase general cognitive load by coercing comprehension into problem solving, thereby engaging mechanisms that play little role in everyday sentence processing (Campbell and Tyler, 2018, Wehbe et al., in prep; Diashek et al., in prep). It is possible that the language comprehension mechanisms that implement linguistic prediction (Shain et al., under review) are relatively less engaged while domain general executive control mechanisms are relatively more engaged during the processing of constructed stimuli presented without context, perhaps suppressing the influence of preceding words on participants' reading behavior. Further investigation is needed in order to explore this hypothesis.

In any case, it is a statistical truism that negative results do not motivate acceptance of the null hypothesis. Thus, it is possible that frequency effects exist in naturalistic reading but are too small to be detected here. Nevertheless, the failure to find frequency effects in large naturalistic data indicates that any such effects are greatly attenuated in the processing of naturalistic texts in comparison to the processing of constructed stimuli, which circumscribes the importance that any such effects might have in driving comprehension effort during typical reading.

6 Conclusion

This paper explored whether effects of word frequency and predictability are distinguishable in naturalistic sentence processing. Despite the size of the combined dataset, results showed no evidence of separable effects in naturalistic reading, contrary to previous findings of separable effects in studies using constructed stimuli. This investigation thus shows no evidence of a distinct, context-independent lexical retrieval mechanism modulated by strength of memory representation (Seidenberg and McClelland, 1989; Coltheart et al., 2001; Harm and Seidenberg, 2004), and instead favors a view in which sentence processing effort is driven by a mechanism that incrementally reallocates resources between competing interpretations, subsuming any effects of raw lexical frequency (Norris, 2006; Levy, 2008; Rasmussen and Schuler, 2018). The discrepancy between constructed and naturalistic experimental settings presents a puzzle for our understanding of the mental processes that underlie human language comprehension, and is perhaps linked to recent evidence that artificially constructed linguistic stimuli can spuriously engage non-linguistic executive mechanisms by increasing general cognitive load as compared to naturalistic settings (Blank and Fedorenko, 2017; Campbell and Tyler, 2018). Further investigation into the precise sources of the discrepancy may shed new light on the interplay between prediction and memory in human sentence processing.

Acknowledgements

The author would like to thank the anonymous reviewers for their helpful comments. This work was supported by National Science Foundation grants #1551313 and #1816891. All views expressed are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Jane Ashby, Keith Rayner, and Charles Clifton. 2005. Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology Section A*, 58(6):1065–1086.

Harald Baayen, Shravan Vasishth, Reinhold Kliegl, and Douglas Bates. 2017. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94(Supplement C):206–234.

R Harald Baayen, Jacolien van Rij, Cecile de Cat, and Simon Wood. 2018. Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In Dirk Speelman, Kris Heylen, and Dirk Geeraerts, editors, *Mixed Effects Regression Models in Linguistics*. Springer, Berlin.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Idan Blank and Evelina Fedorenko. 2017. Domain-general brain regions do not track linguistic input as closely as language-selective regions. *Journal of Neuroscience*, pages 3616–3642.

Andreja Bubic, D Yves Von Cramon, and Ricarda I Schubotz. 2010. Prediction, cognition and the brain. *Frontiers in human neuroscience*, 4:25.

Karen L Campbell and Lorraine K Tyler. 2018. Language-related domain-specific and domain-general systems in the human brain. *Current opinion in behavioral sciences*, 21:132–137.

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.

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

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan):1–30.

Timothy Dozat. 2016. Incorporating Nesterov momentum into Adam. In *ICLR Workshop*.

Susan F Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Stefan L Frank and Rens Bod. 2011. Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological Science*, 22(6):829–834.

Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

Richard Futrell, Edward Gibson, Harry J . Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Tamar H Gollan, Timothy J Slattery, Diane Goldenberg, Eva Van Assche, Wouter Duyck, and Keith Rayner. 2011. Frequency drives lexical access in reading but not in speaking: The frequency-lag hypothesis. *Journal of Experimental Psychology: General*, 140(2):186.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18.

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

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8.

Michael W Harm and Mark S Seidenberg. 2004. Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological review*, 111(3):662.

Uri Hasson and Christopher J Honey. 2012. Future trends in Neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage*, 62(2):1272–1278.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. 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.

Albrecht Werner Inhoff and Keith Rayner. 1986. Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception \& psychophysics*, 40(6):431–439.

David January, John C Trueswell, and Sharon L Thompson-Schill. 2009. Co-localization of Stroop and syntactic ambiguity resolution in Broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience*, 21(12):2434–2444.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Edith Kaan and Tamara Y Swaab. 2002. The brain circuitry of syntactic comprehension. *Trends in cognitive sciences*, 6(8):350–356.

Georg B Keller and Thomas D Mrsic-Flogel. 2018. Predictive Processing: A Canonical Cortical Computation. *Neuron*, 100(2):424–435.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6.

Gina R Kuperberg, Phillip J Holcomb, Tatiana Sitnikova, Douglas Greve, Anders M Dale, and David Caplan. 2003. Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. *Journal of Cognitive Neuroscience*, 15(2):272–293.

M Kutas and S A Hillyard. 1984. Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947):161–163.

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

Maryellen C MacDonald, Neal J Pearlmutter, and Mark S Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676–703.

William D Marslen-Wilson. 1975. Sentence Perception as an Interactive Parallel Process. *Science*, 189(4198):226–228.

Yurii E Nesterov. 1983. A method for solving the convex programming problem with convergence rate O(1/kˆ2). In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547.

Dennis Norris. 2006. The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological review*, 113(2):327.

Jared M Novick, John C Trueswell, and Sharon L Thompson-Schill. 2005. Cognitive control and parsing: Reexamining the role of Brocas area in sentence comprehension. *Cognitive, Affective, \& Behavioral Neuroscience*, 5(3):263–281.

Nathan E Rasmussen and William Schuler. 2018. Left-Corner Parsing With Distributed Associative Memory Produces Surprisal and Locality Effects. *Cognitive science*, 42:1009–1042.

Keith Rayner, Jane Ashby, Alexander Pollatsek, and Erik D Reichle. 2004. The effects of frequency and predictability on eye fixations in reading: Implications for the EZ Reader model. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4):720.

Marten van Schijndel and Tal Linzen. 2018. A Neural Model of Adaptation in Reading. In *EMNLP 2018*, pages 4704–4710.

Marten van Schijndel and William Schuler. 2013. An Analysis of Frequency- and Memory-Based Processing Costs. In *Proceedings of Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the ACL*.

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

Mark S Seidenberg and James L McClelland. 1989. A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523.

Cory Shain. 2018. Deconvolutional time series regression: A technique for modeling temporally diffuse effects. Technical report, The Ohio State University, Columbus.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Computational Linguistics for Linguistic Complexity Workshop*, pages 49–58. Association for Computational Linguistics.

Cory Shain, Marten van Schijndel, and William Schuler. 2018. Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Workshop on Linguistic and Neuro-Cognitive Resources (LREC 2018)*.

Cory Shain and William Schuler. 2018. Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Yosef Singer, Yayoi Teramoto, Ben D B Willmore, Jan W H Schnupp, Andrew J King, and Nicol S Harper. 2018. Sensory cortex is optimized for prediction of future input. *eLife*, 7:e31557.

Nathaniel J Smith and Roger Levy. 2011. Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the 33rd CogSci Conference*.

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

Adrian Staub. 2015. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Adrian Staub and Ashley Benatar. 2013. Individual differences in fixation duration distributions in reading. *Psychonomic Bulletin & Review*, 20(6):1304–1311.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C E Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

Shravan Vasishth, Daniela Mertzen, Lena A Jäger, and Andrew Gelman. 2018. The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 103:151–175.

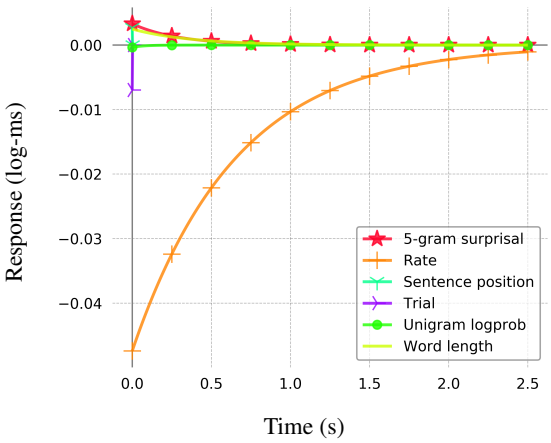

Figure 1: Natural Stories IRFs

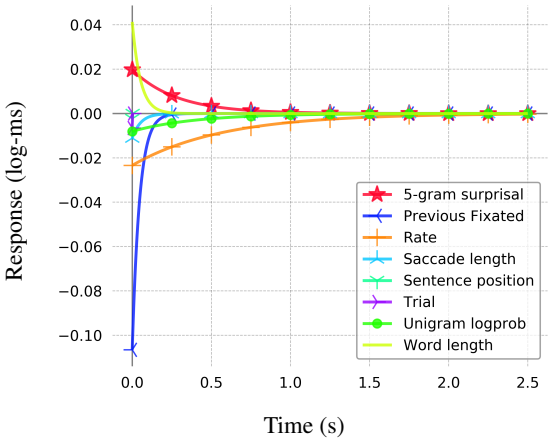

Figure 2: Dundee IRFs

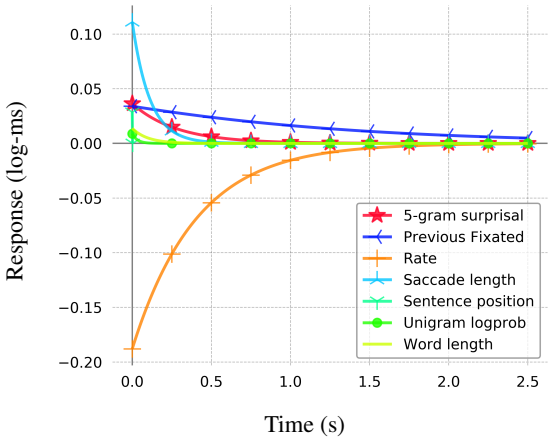

Figure 3: UCL IRFs

A DTSR Implementation

The deconvolutional time series regression (DTSR) models used in this paper were fitted using the code repository released by Shain and Schuler (2018), available at: `https://github.com/coryshain/dtsr`. Models used variational inference to fit the means and variances of independent normal posterior distributions over all model parameters assuming an improper uniform prior. Convolved predictors used the three-parameter ShiftedGamma impulse response function (IRF) kernel:

$$f(x;\alpha,\beta,\delta) = \frac{\beta^{\alpha}(x-\delta)^{\alpha-1}e^{-\beta(x-\delta)}}{\Gamma(\alpha)} \quad (1)$$

Posterior means for the IRF parameters were initialized at $\alpha = 2$, $\beta = 5$, and $\delta = -0.2$, which defines a decreasing IRF with peak centered at $t = 0$ that decays to near-zero within about 1s. Models were fitted using the Adam optimizer (Kingma and Ba, 2014) with Nesterov momentum (Nesterov, 1983; Dozat, 2016), a constant learning rate of 0.01, and minibatches of size 1024. For computational efficiency, histories were truncated at 128 timesteps. Prediction from the network used an exponential moving average of parameter iterates with a decay rate of 0.999, and models were evaluated using *maximum a posteriori* estimates obtained by setting all parameters to their posterior means.[8] Convergence was visually diagnosed.

[8]Since all parameters have independent normal distributions in the variational posterior, the law of large numbers guarantees that samples from the posterior converge in probability to the posterior mean.

B Impulse response shapes

For reference, estimated impulse response shapes by corpus are plotted in Figures 1–3. Plotted curves describe the estimated change in the response t seconds after having observed a unit impulse of each predictor. For example, in the Dundee estimates, observing a word with one standard deviation of *5-gram surprisal* (red curve) is expected to increase reading time by about 0.04 log-ms instantaneously, and by about 0.01 log-ms at a subsequent word observed 0.5s later. Positive IRFs (curves above 0) mean that predictors are estimated to increase reading time (and, by assumption, comprehension difficulty), and negative IRFs (curves below 0) mean that predictors are estimated to decrease reading time. For more detailed psycholinguistic interpretation of IRF estimates like these, see Shain and Schuler (2018).