A large-scale deconvolutional study of predictability and frequency effects in naturalistic reading

Are there separable effects of a word's frequency vs. predictability in human sentence comprehension? Recent work in cognitive science implicates prediction as a major organizing principle in human and animal cognition [1, 23, 13], and psycholinguists have long studied the role of prediction in human sentence processing and its relation to other comprehension mechanisms [18, 15, 17, 27, 9, 19, 16, 5]. 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 [20, 3, 10]. Other theories hold that apparent effects of frequency are underlyingly effects of predictability [19, 16]. A number of studies using constructed stimuli that factorially manipulate frequency and predictability have found separable additive effects of each variable, supporting the position that frequency and predictability index distinguishable influences on lexical processing (see [26] for a review). However, such studies usually use cloze estimates of predictability, which are known to have difficulty differentiating degrees of low contextual probability [24, 25]. Furthermore, while constructed stimuli afford direct control over linguistic variables, results may be influenced by task-specific artifacts and should therefore be complemented by naturalistic studies [4, 11, 22].

This study explores the generalizability of these findings to typical sentence comprehension by searching for separable effects of frequency and predictability during naturalistic reading. Although naturalistic data address the aforementioned concerns about ecological validity, they have their own potential shortcomings that are addressed here in various ways. First, deconvolutional time series regression (DTSR) is used to address the possibility of temporally overlapping response profiles that violate the independence assumptions of linear regression and may therefore confound model interpretation and hypothesis testing [21]. Second, held-out evaluation is used to incorporate model validity directly into statistical tests, avoiding approaches (e.g. likelihood ratio testing) that implicitly evaluate on in-sample data [28]. Third, the natural collinearity between frequency and predictability [4] is addressed through large-scale data, specifically three large naturalistic reading time corpora: Natural Stories (self-paced reading) [7], Dundee (eye-tracking) [14], and UCL (eye-tracking) [6]. The corpora contain over one million data points in total generated by 243 human subjects. Failure to distinguish effects of frequency and predictability would therefore raise doubts about the existence of such a separation in naturalistic sentence comprehension.

Predictability and frequency are operationalized using 5-gram and unigram language models (respectively), each computed using KenLM [12] trained on the Gigaword 3 corpus [8]. Models fit ShiftedGamma impulse response functions [21] to these variables, as well as to the nuisance variable *word length*, along with (eye-tracking only) *saccade length* and an indicator variable for whether the previous word was fixated. Furthermore, to capture trends in the response at different timescales, models contain linear effects for the word's index in the sentence (*sentence position*) and document (*trial*). Following [21], in addition to the intercept, models also fit a convolved intercept (*rate*) designed to capture effects of stimulus timing. The response used in all corpora was log fixation duration (go-past for eye-tracking).[1] Outlier filtering is performed in each corpus following the procedures described in [21]. Approximately half the data in each corpus is used for fitting, with the remaining half reserved for held-out evaluation. Models include by-subject random intercepts as well as by-subject slopes and impulse response parameters for each predictor.[2] Held-out hypothesis testing use a "diamond" ablative structure (*5-gram surprisal* vs. *unigram logprob*) via paired permutation test of the by-item losses on the evaluation set, pooling across all corpora.[3] If predictability and frequency effects are additive, all four comparisons should be significant.

As shown in Table 1, this is not the case. While results show evidence that both frequency and predictability in isolation reliably index processing difficulty (both improve significantly over the baseline), 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 [19, 16] or (2) a more circumscribed role for frequency effects in everyday sentence processing than constructed experiments would suggest, possibly due to task artifacts induced by such experiments [4, 11, 2].

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

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

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

Comparison	*p*-value
5-gram only vs. baseline	**0.0001*****
Unigram only vs. baseline	**0.0001*****
5-gram + Unigram vs. Unigram-only	**0.0001*****
5-gram + Unigram vs. 5-gram-only	0.1440

Table 1: Pooled hypothesis testing results.

	Effect estimate (log-ms)							
Corpus	SentPos	Trial	Rate	WordLen	SacLen	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 2: Effect estimates in log-ms by corpus, computed as the integral of the impulse response over the longest time offset seen in training [21]. Following psycholinguistic convention, unigrams were encoded as log-probabilities while 5-grams were encoded as negative log probabilities (surprisal), resulting in opposite signs. Only one estimate for *sentence position*/*trial* is reported for UCL because sentences were shuffled, rendering *sentence position* and *trial* identical.

References

[1] Andreja Bubic, D Yves Von Cramon, and Ricarda I Schubotz. *Frontiers in human neuroscience*, 2010.

[2] Karen L Campbell and Lorraine K Tyler. *Current opinion in behavioral sciences*, 2018.

[3] Max Coltheart, Kathleen Rastle, Conrad Perry, et al. *Psychological review*, 2001.

[4] Vera Demberg and Frank Keller. *Cognition*, 2008.

[5] Stefan L Frank and Rens Bod. *Psychological Science*, 6 2011.

[6] Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. *Behavior Research Methods*, 2013.

[7] Richard Futrell, Edward Gibson, Harry J . Tily, et al. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 5 2018.

[8] David Graff and Christopher Cieri. English Gigaword LDC2003T05, 2003.

[9] John Hale. In *Proceedings of the Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 2001.

[10] Michael W Harm and Mark S Seidenberg. *Psychological review*, 2004.

[11] Uri Hasson and Christopher J Honey. *NeuroImage*, 2012.

[12] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 8 2013.

[13] Georg B Keller and Thomas D Mrsic-Flogel. *Neuron*, 2018.

[14] Alan Kennedy, James Pynte, and Robin Hill. In *Proceedings of the 12th European conference on eye movement*, 2003.

[15] M Kutas and S A Hillyard. *Nature*, 1984.

[16] Roger Levy. *Cognition*, 2008.

[17] Maryellen C MacDonald, Neal J Pearlmutter, and Mark S Seidenberg. *Psychological Review*, 1994.

[18] William D Marslen-Wilson. *Science*, 1975.

[19] Dennis Norris. *Psychological review*, 2006.

[20] Mark S Seidenberg and James L McClelland. *Psychological review*, 1989.

[21] Cory Shain and William Schuler. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[22] Cory Shain, Marten van Schijndel, and William Schuler. In *Workshop on Linguistic and Neuro-Cognitive Resources (LREC 2018)*, 2018.

[23] Yosef Singer, Yayoi Teramoto, Ben D B Willmore, et al. *eLife*, 2018.

[24] Nathaniel J Smith and Roger Levy. In *Proceedings of the 33rd CogSci Conference*, 2011.

[25] Nathaniel J Smith and Roger Levy. *Cognition*, 2013.

[26] Adrian Staub. *Language and Linguistics Compass*, 2015.

[27] Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C E Sedivy. *Science*, 1995.

[28] Samuel S Wilks. *The Annals of Mathematical Statistics*, 1938.