**Low-level language statistics affect reading times independently of surprisal**

Surprisal theory has provided a successful unifying framework for understanding a wide range of phenomena in sentence processing (Hale, 2001; Levy, 2008). This theory states that the probability of a word given its preceding context $p(w_n|w_1^{n-1})$ is a *causal bottleneck* in linking linguistic properties to reading times; e.g., lower-level statistics such as the frequency of a word or construction only affect reading times via influencing $p(w_n|w_1^{n-1})$, which itself affects reading times. Problematically for this strong claim, it has been well documented empirically that one low-level statistic, word frequency, affects reading times independently of surprisal. Such results may suggest a broad role for low-level statistics in processing, independent of surprisal, or word frequency may be a special case. Here, we present what we believe is the first clear evidence that a more complex low-level statistic, bigram probability $p(w_n|w_{n-1})$, also affects processing independently of surprisal, suggesting a broad, independent role of low-level statistics in processing.

Some prior work has reported effects of bigram probability independent of surprisal (Demberg & Keller, 2008; Fossum & Levy, 2012). However, this work may in fact be compatible with surprisal theory because of properties of how the surprisal and bigram probabilities were calculated. A surprisal probability $p(w_n|w_1^{n-1})$ and a bigram probability $p(w_n|w_{n-1})$ are typically estimated (imperfectly) by computational language models. Often, estimation error across language models is somewhat independent, and thus taking a weighted average (interpolation) of two language models $p_{interp}(w_n|w_1^{n-1}) = \gamma p_1(w_n|w_1^{n-1}) + (1-\gamma)p_2(w_n|w_1^{n-1})$ can often yield a better language model than either individually. This is especially true if the two language models are different classes of model, as was the case for the prior results. Thus, simultaneous effects of surprisal and bigram probability in prior work may actually result from a regression model implicitly interpolating the surprisal and bigram language model probabilities to form a better estimate of surprisal.

We go beyond prior work by (a) estimating surprisal using a model class that captures frequency and bigram information and (b) demonstrating directly that interpolating this estimate with frequency and bigram probabilities does not improve it. We then analyze gaze durations in an eye movement corpus as a function of these surprisal, frequency, and bigram probability predictors.

**Perplexity analysis**. We estimated surprisal, bigram probability, and frequency using the Google One Billion Word Benchmark corpus (Chelba et al., 2013). Our language model for surprisal was an interpolation of a smoothed 5-gram model and a Long Short-Term Memory (LSTM) network (Jozefowicz et al., 2016). To ensure generality, we tested two versions of the interpolation weight, one that optimized perplexity (*fitted*), and one that equally weighted the two component models (*balanced*). Frequency and bigram probability were estimated by smoothed n-gram models.

For each pair of surprisal model and frequency/bigram model, we used grid search to compute the interpolation weights that optimized perplexity (ability of the language model to predict words). Results showed that interpolating with lower-order statistics never improved the surprisal model.

**Gaze duration analysis**. We analyzed word gaze durations from the Dundee eye movement corpus (Kennedy et al., 2003) using two mixed-effects regressions, for the Optimized and Balanced surprisals. Predictors of interest were log frequency and bigram probability of $w_n$ and $w_{n-1}$ and we included random slopes of these by subject. Fixed effect covariates (modeled on Smith & Levy, 2013) included surprisal, length, and the frequency–length interaction for $w_n$ and $w_{n-1}$, prior word fixation, and text position. Results (Table 1) showed significant effects of frequency and bigram probability of $w_n$ and $w_{n-1}$ on gaze durations (though $w_{n-1}$ bigram probability was unclear).

**Conclusion**. These results demonstrate that two low-level statistics, frequency and bigram probability, have effects on processing that cannot be explained by surprisal theory. They motivate research into new generalizations of surprisal (such as noisy-channel surprisal, Futrell & Levy, 2017) that can also explain why local statistical information should have an outsized effect.

| Language Model | Predictor | $\hat{\beta}$ | $\chi^2_1$ | |
|---|---|---|---|---|
| Fitted | log frequency $w_n$ | $-11.58$ | $9.23$ | $p < 0.01$ |
| Fitted | log bigram probability $w_n$ | $-1.49$ | $4.52$ | $p < 0.05$ |
| Fitted | log frequency $w_{n-1}$ | $-2.16$ | $8.01$ | $p < 0.01$ |
| Fitted | log bigram probability $w_{n-1}$ | $-0.74$ | $2.93$ | $p = 0.08$ |
| Balanced | log frequency $w_n$ | $-11.61$ | $9.25$ | $p < 0.01$ |
| Balanced | log bigram probability $w_n$ | $-1.38$ | $3.88$ | $p < 0.05$ |
| Balanced | log frequency $w_{n-1}$ | $-2.29$ | $8.69$ | $p < 0.01$ |
| Balanced | log bigram probability $w_{n-1}$ | $-0.39$ | $0.81$ | $p = 0.36$ |

Table 1: Results of linear mixed-effects regressions predicting gaze durations when using the Fitted surprisal model (top 4 rows) and Balanced surprisal model (bottom 4 rows): estimated coefficients and results of likelihood ratio tests for frequency and bigram probability predictors.

## References


Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., & Robinson, T. (2013). One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

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

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics* (pp. 61–69).

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics* (pp. 688–698).

Hale, J. (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* (pp. 1–8).

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Kennedy, A., Hill, R., & Pynte, J. (2003). The dundee corpus. In *Proceedings of the 12th european conference on eye movement.*

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

Smith, N., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.