

Global sentence properties rather than local probabilities drive wordform selection

Cassandra L. Jacobs^{1,2} & Maryellen C. MacDonald¹

1: University of Wisconsin-Madison; 2: University at Buffalo

cxjacobs@buffalo.edu

What constrains lexical choice, such as the decision to produce a full noun phrase (e.g., “the sleeping cat”) versus a “reduced” referring expression, like pronoun (“they”)? Information theoretic accounts of reduction state that the probability or improbability of a word in context constrains its form (Aylett & Turk, 2004; Jaeger, 2010; Mahowald et al., 2013). Others have argued for a more global, message-level approach in which broad contextual factors like discourse register affect lexical selection (Kapatsinski, 2009). In this work, we ask a related question: how much do **global context** (e.g., discourse register) and **local surprisal** (an information theoretic metric) contribute to production preferences for short (*math*) and long (*mathematics*) English wordforms in a sentence rating task (Experiment 1) and naturally occurring corpora (Experiment 2).

Experiment 1: Predicting choice in a sentence rating task. We selected 38 pairs of related words (76 total) from Mahowald et al. (2013) and designed 10 sentences for each pair. Half (5) of the sentences were designed to encourage the short (*math*) or long form (*mathematics*) of a pair (e.g., “Professor Vector was awarded for his notable advancements in research for both ___ and the sciences.”). 91 American English speakers rated which of the two lexical alternatives best fit in the blank (one sentence per pair). We built mixed effects models of participant decisions showed that the higher the **median log word frequency** of all other words in the sentence, the more the short form was preferred ($\beta=-0.39$, $Z=-4.81$, $p < .001$). **Surprisal** calculated from a state-of-the-art neural language model from the upstream sentence context (Ng et al., 2019) did not significantly predict word forms ($\beta=0.09$, $Z=0.83$, $p=n.s.$). Combining both factors worsened model fit relative to the frequency-only model. This suggests that one index of global sentence context (frequency) more strongly predicts lexical choice than local context (surprisal).

To test how much **latent contextual factors alone** predict preferences for short and long words, we built 38 separate LASSO logistic regression models to predict participants’ preferences for short or long words using only a vector (embedding) representation of each sentence taken from a **masked language model** (Liu et al., 2019), from which we hid the critical word. Each model was trained on all ratings for 370 sentences with no covariates and produced one probability value corresponding to the probability of receiving a “long” rating. for the remaining 10 sentences. Ratings for long versus short forms for 37 of the other pairs **generalized** to preferences for the 38th pair (Table 1, Figure 1). This shows that **general contextual factors** predict word length, supporting register accounts.

Experiment 2: Predicting choice in natural corpora. We sought to confirm that participants were using linguistic knowledge rather than picking up on unintended patterns in our stimuli. We extracted up to 200 randomly selected sentences from COCA (years 1990-2015; Davies, 2008) for each of the 76 unique words, which we turned into vectors in the same manner as in Experiment 1, and applied the trained classifiers from Experiment 1 to obtain a probability that the long word from a pair would appear in that sentence. We **replicated the effect of context** and found that participants’ length preferences generalized to naturally occurring sentences (Table 2, Figure 2), such that if participants would have preferred the longer in a sentence, that sentence was more likely to actually contain the longer form. This provides additional support in favor of the global contextual hypothesis.

Conclusion. This work demonstrates that global sentence-level properties, such as the frequencies of surrounding words and latent properties of the entire sentence predict participants’ preferences for what words are appropriate in a given context. Critically, the surprisal of a word in context did not meaningfully predict participants’ preferences. The results suggest that participants’ preference for **word length varies with aspects of the global context, consistent with register influences on lexical selection**. These results motivate more exploration into how information theoretic accounts of lexical selection should be adjusted to permit message-level utterance planning.

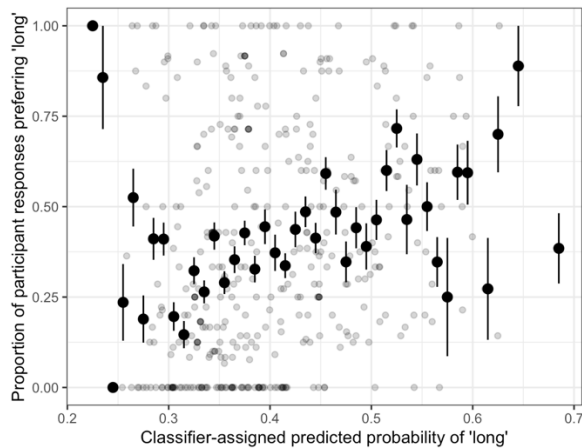


Figure 1: By-item classifier model predicted probabilities and ratings. Binned for visualization.

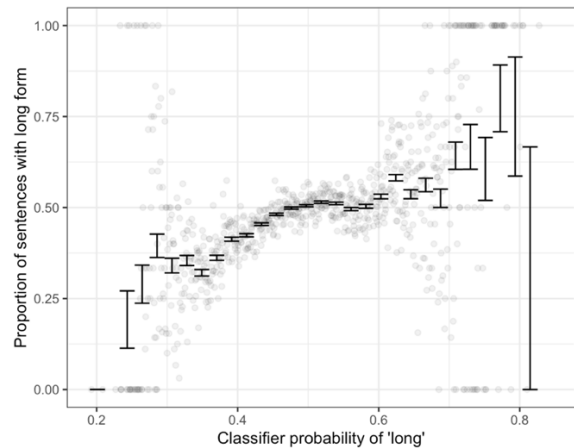


Figure 2: Correlation between predicted participant ratings from classifier trained on RoBERTa embeddings of COCA sentences and corpus probabilities. Binned for visualization.

Name	β	SE	Z	p
Intercept	-0.62	0.27	-2.34	< .05
Classifier predicted probability	0.55	0.11	4.84	< .001

Table 1: Model predicting participants' word form preferences using out-of-sample word embedding-based model predicted score

Name	β	SE	Z	p
Intercept	-0.07	0.05	-1.54	n.s.
Classifier predicted probability	0.24	0.08	2.90	< .01

Table 2: Model predicting word form in COCA from participant ratings

References:

- Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47, 31-56.
- Davies, M. (2008). The corpus of contemporary American English (COCA): 560 million words, 1990-present.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61, 23-62.
- Kapatsinski, V. (2009). Adversative conjunction choice in Russian (no, da, odnako): Semantic and syntactic influences on lexical selection. *Language Variation and Change*, 21, 157-173.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.
- Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126, 313-318.
- Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019, August). Facebook FAIR's WMT19 News Translation Task Submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)* (pp. 314-319).