High frequency derived words have low semantic transparency mostly only if they are polysemous

Martha Booker Johnson The Ohio State University
Micha Elsner The Ohio State University
Andrea D. Sims The Ohio State University

The development of word vectors as an implementation of distributional semantics (Boleda, 2020, inter alia) offers new tools for quantitatively testing old ideas about morphology. Since Bybee (1985), an often-repeated claim is that a strong relationship holds between a derived lexeme's token frequency and its semantic relationship to its base. Specifically, high frequency is posited to facilitate low semantic transparency as a function of lexical storage (Baayen, 1993; Bybee, 1985). Yet surprisingly little work has tested this using a quantitative measure of semantic transparency. Closest is Hay (2001), who treats semantic transparency as binary. We start from the observation that polysemy complicates base-derivative relations (e.g. Lapesa et al., 2018; Salvadori & Huyghe, 2023). An open question thus has to do with the role that polysemy plays in the relationship between derivative frequency and semantic transparency, if any. We use word vectors to test for a correlation between semantic transparency and derivative frequency in English, examining the role of polysemy.

We present three analyses. First, using a large dataset we show that the simple claim of an inverse relationship between derivative frequency and semantic transparency (operationalized as cosine similarity) is not supported, contrary to received wisdom. Second, using a subset of the data we show that the expected relationship *can* be detected, but only when interactions between frequency and polysemy are considered. Finally, we validate this result by showing that similar polysemy effects also emerge in human judgments of the semantic relatedness of bases and derivatives. Specifically, high polysemy derivatives exhibit an inverse relationship between derivative frequency and semantic transparency but low polysemy derivatives do not.

In short, polysemy mediates the relationship between frequency and semantic transparency, a fact that has not been sufficiently recognized in previous work.

1. No simple correlation between frequency and semantic transparency

To test for a correlation between derivative frequency and semantic transparency in the English lexicon, we started with 10,465 derived English lexemes from Sims & Parker (2015), which correspond to all of the lexemes in CELEX (Baayen et al., 1995) that end in one of 54 English derivational suffixes. We extracted these lexemes' bases from CELEX's morphological analysis. Lemma frequency for each derivative and base word was calculated from the training set of the Tensorflow Wiki40b dataset (Guo et al., 2020), which provides English Wikipedia data cleaned of extraneous text. We lemmatized and part-of-speech (POS) tagged the dataset using CoreNLP (Manning et al., 2014) and then calculated token frequency for each lemma. For all models we converted frequency counts to log instances per million words of corpus (log ipm) since word frequencies are Zipfian. Base-derivative pairs in which either lemma had fewer than 8 tokens ($= 0.1$ ipm) were removed because word vectors are typically unstable for low-frequency items. Suffixes with fewer than 10 example pairs were then also dropped. This resulted in a dataset containing 3,286 base-derivative pairs for 34 suffixes.

We operationalize semantic transparency as the cosine similarity of a derived lexeme's vector to its base lexeme's vector. For each base and derived lexeme, we retrieved its 300-dimensional vector from Fares et al.'s (2017) lemmatized English model that was trained on

the English Wikipedia dump of February 2017. We then calculated cosine similarity (ranging between 0 and 1, with higher values indicating greater semantic transparency) for each base-derivative pair. (Cosine similarity was chosen as a measure in order to maximize comparability to the task in Analysis 3, which asked participants to compare the similarity of base and derivative forms.)

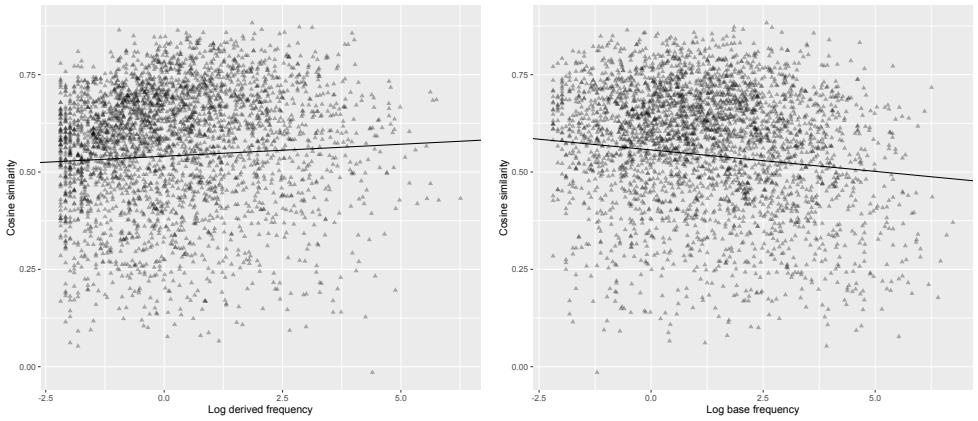

Figure 1: (Lack of) correlation between derived frequency (in log ipm) and cosine similarity (left panel), or base frequency (in log ipm) and cosine similarity (right panel)

As shown in Figure 1, the data are widely dispersed. We constructed a mixed effects regression model that predicted cosine similarity from derived frequency, with suffix and base as random intercepts. A negative relationship between derivative frequency and cosine similarity was expected. Results show frequency was significantly *positively* correlated with cosine similarity ($\beta = 0.006$, $t(3277) = 3.766$, $p < 0.001$) but had an extremely low marginal $R^2 = 0.0017$, indicating that word frequency accounts for almost none of the variance. Thus, no simple, robust correlation between low semantic transparency and high frequency is observed. (A model with base frequency as a predictor produced a similarly weak pattern in the opposite direction.)

2. A correlation exists, but predominantly for highly polysemous derived words

Analysis 1 suggested that the null result was caused by uncontrolled variables, with polysemy as the main suspect. To investigate this we used 109 base-derivative pairs (a subset of the Analysis 1 data) that had been experimental stimuli for a ratings task (McKenzie, 2019). Derivative frequency and base frequency were strongly correlated, so to use both as predictors in the model we residualized base frequency on derived frequency. The number of senses of the derivative and of the base (our measures of polysemy) was calculated as the number of senses listed in the online Oxford English Dictionary (oed.com). A final, stepped-down mixed effects regression model had the following fixed effects: derived frequency, squared derived frequency, residualized base frequency, squared residualized base frequency, and derived number of senses. There was one two-way interaction: residualized base frequency*derived number of senses. There was a random intercept for affix. All factors were centered or sum-contrasted, as appropriate.

The left panel of Figure 2 visualizes the main effect (and quadratic) relationship between derivative frequency and cosine similarity, in the expected direction (i.e. a negative correlation). Even more interesting is the two-way interaction shown in the right panel. Since base frequency was residualized on derived frequency, an x-axis value of 0 represents a base that is exactly as frequent as would be expected given the frequency of the derived word. Negative values indicate base-derivative pairs in which base frequency is lower than expected (or equivalently, derived frequency is higher). The values for derived number of senses are the mean

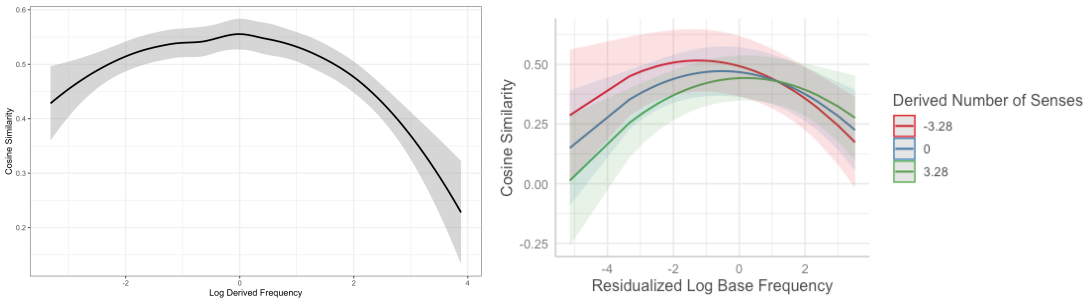

Figure 2: Model-predicted values for cosine similarity based on derived frequency (left panel) and interaction between residualized base frequency and derived number of senses (right panel)

(blue) and one standard deviation above (green) and below (red) the mean. As shown, among derivatives with higher-than-expected frequency, lower cosine similarity values are observed for those that are also highly polysemous. The correlation between low semantic transparency and high frequency is thus more strongly a property of highly polysemous derived words.

3. A similar pattern emerges in similarity judgments

Finally, as a check on whether the word vectors are adequately capturing human intuitions about semantic relatedness of base-derivative pairs, we reanalyzed data from McKenzie (2019), which asked 24 native speakers of English to provide semantic similarity judgments for the same 109 pairs used for Analysis 2. Participants responded to the prompt "How similar is the meaning of the word [DERIVED] to the meaning of the word [BASE]?" using a continuous scale. The fixed effects in the final, stepped-down model were derived frequency, residualized base frequency, derived number of senses, and base number of senses. The final model also included three two-way interactions — derived frequency*derived number of senses, residualized base frequency*derived number of senses, and derived number of senses*base number of senses — and random intercepts for participant and word, with affix as a grouping factor for word. All factors were centered or sum-contrasted, as appropriate.

The relationship between derived frequency, derived number of senses and base number of senses is visualized in Figure 3. For words with few derived senses (red line), there is no change in response based on derived frequency. For derived words with average and above average number of senses (blue and green lines), however, participant similarity judgments decrease as derived frequency increases, with a steeper slope for words with more senses. Thus, similarly to what was observed with cosine similarity, a negative relationship between derivative frequency and semantic transparency is characteristic only of polysemous derivatives. Additionally, an interaction between base polysemy and derivative polysemy is observed. For low polysemy derivatives (red line), as the polysemy of the base increases, semantic transparency judgments decrease. Thus, polysemy of both the derivative and the base affects judgments.

Vector models of word-formation have proliferated recently, showing that the distributional semantic approach can be profitably applied to a range of morphological questions. Our study contributes to this line of research. Specifically, we show that frequent claims suggesting that high word token frequency is straightforwardly correlated with low semantic transparency do not hold in English. Instead, the relationship is crucially mediated by polysemy. The most semantically opaque derivatives have high frequency **and** are highly polysemous. Thus, despite being received wisdom, the relationship between frequency and semantic transparency (in English) is more complex than previously understood. Ongoing work includes implementation

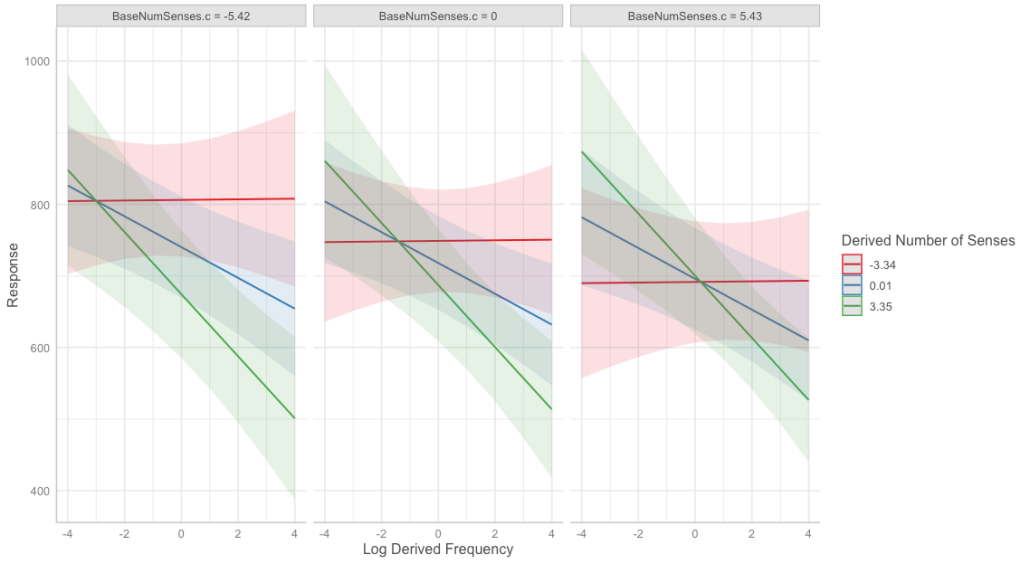

Figure 3: Model-predicted values for similarity judgments, with derived frequency (x-axis), derived number of senses (series; red = low, green = high), and base number of senses (panels)

of Marelli & Baroni's (2015) measure of semantic transparency, whcih calculates compositional vectors, to take account of the affix's semantic contribution, as well as expanding the dataset (Analysis 2) to include polysemy information for a larger number of base-derivative pairs.

References

Baayen, R. H. 1993. On frequency, transparency, and productivity. In G. Booij & J. van Marle (eds.), *Yearbook of morphology 1992*, 181–208. Kluwer.

Baayen, R. H., R. Piepenbrock & L. Gulikers. 1995. The CELEX Lexical Database (CD-ROM).

Boleda, G. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics* 6. 213–234.

Bybee, J. 1985. *Morphology: A study of the relation between meaning and form.* John Benjamins.

Fares, M., A. Kutuzov, S. Oepen & E. Velldal. 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa,* 271–276.

Guo, M., Z. Dai, D. Vrandecic & R. Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. *Proceedings of LREC 2020: 12th International Conference on Language Resources and Evaluation* 2440–2452.

Hay, J. 2001. Lexical frequency in morphology: Is everything relative? *Linguistics* 39. 1041–1070.

Lapesa, G., L. Kawaletz, I. Plag, M. Andreou, M. Kisselew & S. Padó. 2018. Disambiguation of newly derived nominalizations in context: A Distributional Semantics approach. *Word Structure* 11. 277–312.

Manning, C., M. Surdeanu, J. Bauer, J. Finkel, S. Bethard & D. McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* 55–60.

Marelli, M. & M. Baroni. 2015. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review* 122. 485–515.

McKenzie, M. 2019. Effects of relative frequency on morphological processing in Russian and English. BA thesis. The Ohio State University.

Salvadori, J. & R. Huyghe. 2023. Affix polyfunctionality in French deverbal nominalizations. *Morphology* 33. 1–39.

Sims, A. & J. Parker. 2015. Lexical processing and affix ordering: Cross-linguistic predictions. *Morphology* 25. 143–182.