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

Martha Booker Johnson, Micha Elsner, & Andrea D. Sims

The Ohio State University, USA

13 September 2023

Roadmap

1. Introduction
2. Analysis 1: No simple correlation between frequency and semantic transparency
3. Analysis 2: A correlation exists, but predominantly for highly polysemous derived words
4. Analysis 3: A similar pattern emerges in similarity judgments
5. Discussion

Table of Contents

Semantic transparency as intuition

- *busyness* $=$ *busy* $+$ *ness*
- *business* $\neq$ *busy* $+$ *ness*

Polysemy

- As one example, at least one meaning of the derived word *movement* is related to the verb *move*; this neglects, however, that *movement* can also refer to a *political movement* or a *movement of a concerto*.
- These other meanings of the whole derived word are not (intuitively) highly related to the verb *move*.

Semantic transparency and frequency

- Word vectors (Boleda, 2020, inter alia) offer new tools for quantitatively testing old ideas about morphology.

Semantic transparency and frequency

- Word vectors (Boleda, 2020, inter alia) offer new tools for quantitatively testing old ideas about morphology.
- Since Bybee (1985), an often-repeated claim (e.g., most recently, Bobkova and Montermini, 2023) 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).

Semantic transparency and frequency

- Word vectors (Boleda, 2020, inter alia) offer new tools for quantitatively testing old ideas about morphology.
- Since Bybee (1985), an often-repeated claim (e.g., most recently, Bobkova and Montermini, 2023) 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).
- Recent work has used word vectors to quantitatively measure the semantics of derivational affixes (e.g. Guzmán Naranjo and Bonami, 2023; Günther et al., 2019; Huyghe and Wauquier, 2021; Kotowski and Schäfer, 2023; Varvara et al., 2021), but doesn't test the claim about frequency.

Semantic transparency and frequency

- Word vectors (Boleda, 2020, inter alia) offer new tools for quantitatively testing old ideas about morphology.
- Since Bybee (1985), an often-repeated claim (e.g., most recently, Bobkova and Montermini, 2023) 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).
- Recent work has used word vectors to quantitatively measure the semantics of derivational affixes (e.g. Guzmán Naranjo and Bonami, 2023; Günther et al., 2019; Huyghe and Wauquier, 2021; Kotowski and Schäfer, 2023; Varvara et al., 2021), but doesn't test the claim about frequency.
- We start from the observation that polysemy complicates base-derivative relations (e.g. Lapesa et al., 2018; Salvadori and Huyghe, 2023).

Questions

- Do predictions that higher derived word frequency corresponds to lower semantic transparency hold when semantic transparency is quantified and calculated for thousands of derived/base word pairs in a language?
- What is the relationship between polysemy, lexical frequency, and semantic transparency?
- How do quantitative measures (specifically cosine similarity) of derived/base pairs correspond (and not correspond) to human judgments of derived/base pair semantic similarity?

Three analyses

1. Large dataset: simple claim of an inverse relationship between derivative frequency and semantic transparency (operationalized as cosine similarity) is not supported

Three analyses

1. Large dataset: simple claim of an inverse relationship between derivative frequency and semantic transparency (operationalized as cosine similarity) is not supported
2. Large dataset with polysemy: derived and base frequency interact in important ways with derived and base word polysemy

Three Analyses

1. Large dataset: simple claim of an inverse relationship between derivative frequency and semantic transparency (operationalized as cosine similarity) is not supported
2. Large dataset with polysemy: derived and base frequency interact in important ways with derived and base word polysemy
3. Human judgments on small dataset: similar polysemy effects emerge

Table of Contents

Methods

- 10,465 derived English lexemes from Sims and Parker (2015), which correspond to all of the lexemes in CELEX (Baayen et al., 1995) that end in one of 54 English derivational suffixes.
- Lemma frequency calculated from training set of the Tensorflow Wiki40b dataset (Guo et al., 2020)
- Data lemmatized and part-of-speech (POS) tagged using CoreNLP (Manning et al., 2014)

Methods cont'd

- Converted frequency counts to log instances per million words of corpus (log ipm)
- Base-derivative pairs in which either lemma had fewer than 8 tokens (= 0.1 ipm) were removed
- Suffixes with fewer than 10 example pairs were then also dropped
- Final dataset: 3,231 base-derivative pairs for 38 suffixes

Analysis 1 & 2: Suffixes

-able	-ate	-ess	-ious	-less	-ry
-age	-ation	-ful	-ish	-ly	-ship
-al	-en	-ian	-ism	-ment	-y
-an	-ence	-ic	-ist	-ness	
-ance	-ent	-ier	-ity	-or	
-ant	-er	-ify	-ive	-ory	
-ary	-ery	-ion	-ize	-ous	

Table: Suffixes used in analyses 1 and 2.

Methods: cosine similarity

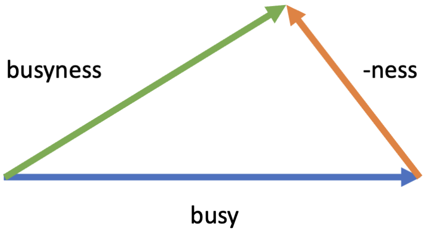


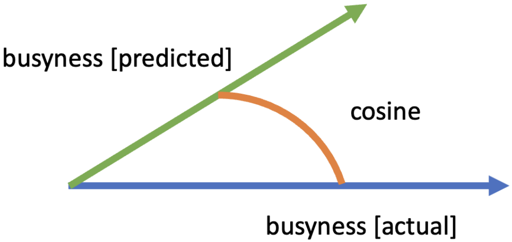
Figure: Toy example of cosine similarity.

- Semantic transparency operationalized as cosine similarity
- Calculated predicted derived vectors by adding affix vector to base vector
- Then compared the predicted to actual derived vector using cosine similarity

Methods: implementational details

- Pre-trained 300-dimensional vectors from Fares et al.'s (2017) lemmatized English model that was trained on the English Wikipedia dump of February 2017
- Calculated affix "meaning" (vector) as average cosine of derived and base vectors
    - Removed target lemma from dataset
    - For remaining words with target affix, subtracted base vectors from derived vectors
    - Took the mean
- Added affix vector to base vector to form predicted vector
- Calculated cosine similarity (= 1 – cosine) of predicted derived word vector and actual derived word vector
- Values closer to 1 mean that the predicted vector is more accurate and, therefore, more transparent

Methods: implementational details cont'd



Model

- Fixed effects in final, stepped-down mixed effects regression model:
  - Derived frequency & derived frequency squared
  - Base frequency & base frequency squared
  - Derived frequency*base frequency

Results

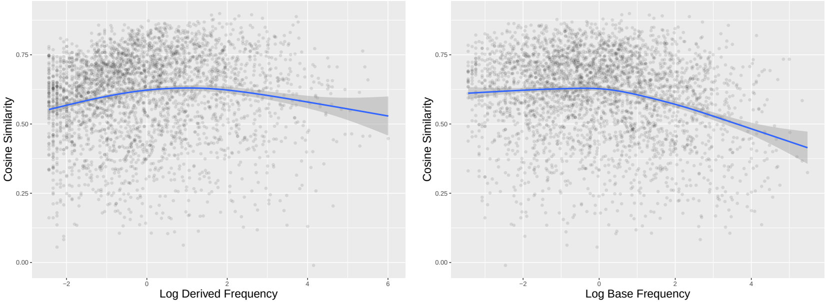


Figure: Raw data with fit lines. Derived frequency (left panel) and base frequency (right panel) are on the x-axes and cosine similarity is on the y-axis.

Results cont'd

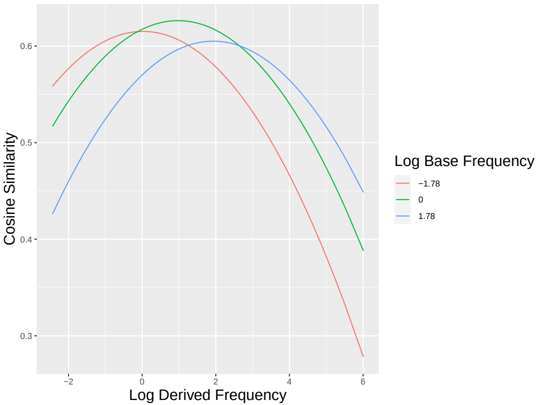


Figure: Relationship between derived frequency (log ipm), base frequency (log ipm), and cosine similarity.

Results cont'd

- Expected pattern of an inverse relationship between derived frequency and semantic transparency holds only for highly frequent words
- More derived words with low frequency bases conform to this expected pattern than derived words with higher frequency bases
- BUT for majority of words on left half of graph derived frequency goes up and cosine similarity goes up
- There is some truth to the conventional story but mostly for high frequency derived words and especially those with infrequent bases
- Conclusion: previous observations are about words on the right edge

Table of Contents

Methods

- Used same data set as analysis 1
- The number of senses of the derivative and of the base was drawn from WordNet (Princeton University, 2010)

Model

Fixed effects in final, stepped-down mixed effects regression model:

- Derived frequency & derived frequency squared
- Base frequency & base frequency squared
- Derived number of senses
- Base number of senses
- Three-way interactions: derived frequency*base frequency*derived senses and base frequency*derived senses*base senses. All two-way interactions were entailed by three-way interactions

First interaction

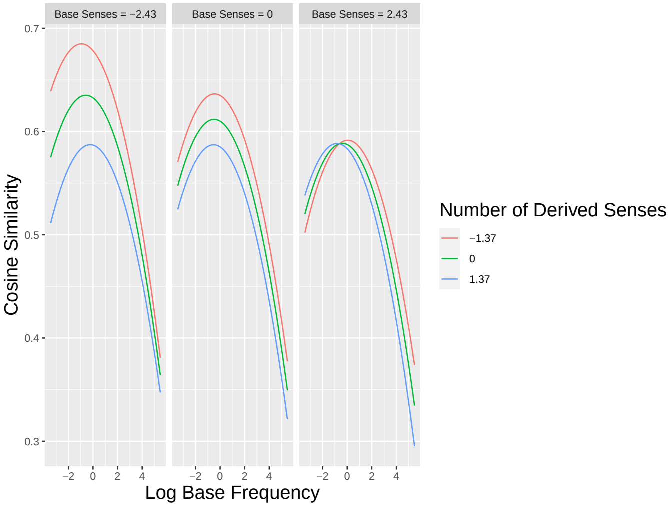


Figure: Model-predicted values for cosine similarity based on base frequency, derived number of senses, and base number of senses.

First interaction cont'd

- When base polysemy is high, derived polysemy doesn't really matter
- Higher cosine similarity depends on low polysemy for both base and derived words
- Semantic transparency is thus strongly influenced by polysemy of both the derivative and the base
- Frequency alone is not enough to account for differences in semantic transparency

Second interaction

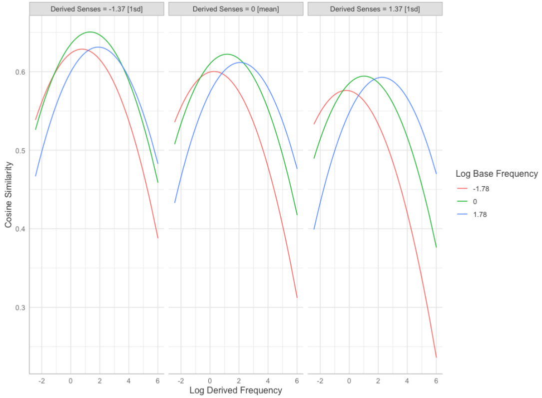


Figure: Model-predicted values for cosine similarity based on derived frequency, base frequency, and derived number of senses.

Second interaction cont'd

- More derived senses -> less cosine similarity
- Base frequency makes a greater difference w/ more senses
- Steeper decrease in semantic transparency as derived frequency increases for *polysemous* derived words

Table of Contents

Methods

- Data from McKenzie (2019)
- 24 native speakers of English provided semantic similarity judgments for 109 derived-base pairs
- They responded to prompt "How similar is the meaning of the word [derived] to the meaning of the word [base]?" using a continuous scale

Suffixes

-ive	-able	-ic	-ess	-ist
-ment	-ate	-ity	-ness	-ism

Table: Suffixes used in analysis 3.

Methods cont'd

- Number of senses of the derivative and of the base calculated as the number of senses listed in the online Oxford English Dictionary (oed.com)
- Residualized base frequency on derived frequency because of excessive collinearity in the model

Methods cont'd

Fixed effects in the final, stepped-down model:

- Derived frequency
- Residualized base frequency
- Derived number of senses
- Base number of senses
- 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

Results

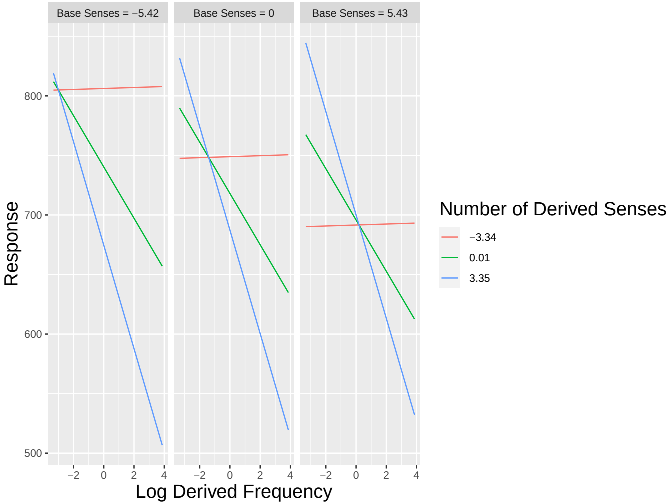


Figure: Model-predicted values for similarity judgments, with derived frequency (x-axis), derived number of senses (colors), and base number of senses (panels).

Results cont'd

- Similarly to what was observed with cosine similarity, a negative relationship between derivative frequency and semantic transparency is characteristic only of polysemous derivatives
- An interaction between base polysemy and derivative polysemy is observed
- Polysemy of both the derivative and the base affects human judgments, so the polysemy results are not an artifact of the method of cosine similarity

Table of Contents

Returning to semantic transparency and frequency

- Is Bybee's (1985) claim that derived frequency negatively correlates with semantic transparency correct?
- There's a strong negative effect of frequency primarily for polysemous derived words
- But, high frequency derived words have high transparency, if they also have low base and derived polysemy

Actual effect or methodological limitation?

- Is this effect really limited to derived-base pairs with low polysemy, or is this a methodological limitation?
- Cosine similarity picks up all the meanings of lemmas, which could alter the outcomes

Semantic transparency

- How we define and implement semantic transparency depends on our goals
- Semantic transparency has often been used to investigate lexical storage
- To use semantic transparency for this, we need a model of how we think polysemy affects lexical storage
- Very large quantitative datasets force us to confront questions that are much easier to avoid when working with hand-chosen examples

Conclusion

- We show that frequent claims suggesting that high word token frequency is straightforwardly correlated with low semantic transparency do not hold in English
- The relationship is crucially mediated by polysemy. The most semantically opaque derivatives have high frequency **and** are highly polysemous
- The relationship between frequency and semantic transparency (in English) is more complex than previously understood

References

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

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

Bobkova, N. and Montermini, F. (2023). A quantitative approach to doublets in Russian denominal adjective constructions. *Word Structure*, 16(1):63–86.

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

Günther, F., Smolka, E., and Marelli, M. (2019). 'Understanding' differs between English and German: Capturing systematic language differences of complex words. *Cortex*, 116:168–175.

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

Guzmán Naranjo, M. and Bonami, O. (2023). A distributional assessment of rivalry in word formation. *Word Structure*, 16:87–114.

Huyghe, R. and Wauquier, M. (2021). Distributional semantics insights on agentive suffix rivalry in french. *Word Structure*, 14:354–391.

Kotowski, S. and Schäfer, M. (2023). Semantic relatedness across base verbs and derivatives: Quantitative and distributional analyses of English *out*-prefixation. In Kotowski, S. and Plag, I., editors, *The semantics of derivational morphology: Theory, methods, evidence*, pages 237–247. De Gruyter, Berlin.

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

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

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

Princeton University (2010). About WordNet.

Salvadori, J. and Huyghe, R. (2023). Affix polyfunctionality in French deverbal nominalizations. *Morphology*, 33:1–39.

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

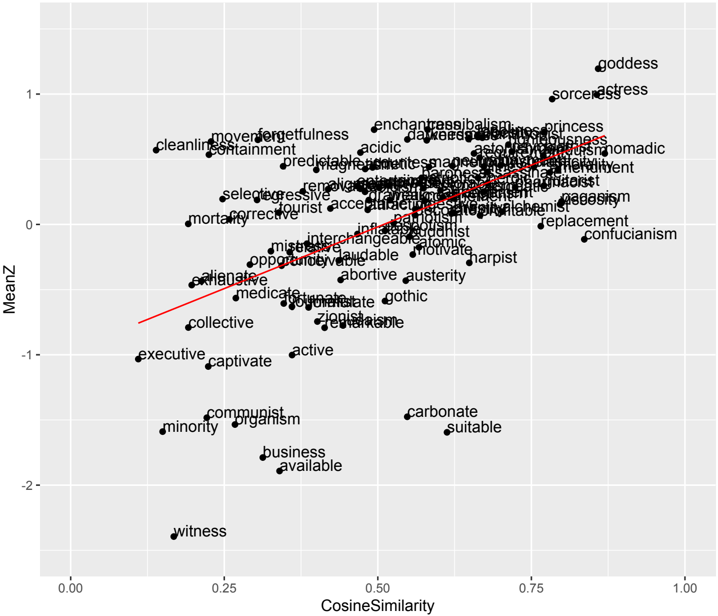
Varvara, R., Lapesa, G., and Padó, S. (2021). Grounding semantic transparency in context: A distributional semantic study on German event nominalizations. *Morphology*, pages 409–446.

Acknowledgments and contact

Thanks to Michelle McKenzie for the use of her data!

Contact: johnson.6713@osu.edu, sims.120@osu.edu
OSF directory: `https://osf.io/q2jsr/`

Some data to consider



Analysis 1: Model results

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.62	0.01119	43.11	55.18	< 0.001
DrvFreq.c	0.01836	0.00197	3224	9.339	< 0.001
DrvFreq.c^2	-0.00941	0.00083	3207	-11.27	< 0.001
BaseFreq.c	-0.01267	0.00163	3224	-7.779	< 0.001
BaseFreq.c^2	-0.00782	0.00069	3211	-11.317	< 0.001
DrvFreq.c*BaseFreq.c	0.01008	0.00099	3215	10.195	< 0.001

Table: Model output for analysis 1.

Analysis 2: model results

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.61	0.01064	44.78	57.359	< 0.001
DrvFreq	0.0206	0.00199	3216	10.388	< 0.001
DrvFreq^2	-0.0088	0.00091	3200	-9.602	< 0.001
BaseFreq	-0.0072	0.00179	3214	-4.034	< 0.001
BaseFreq^2	-0.0074	0.00075	3203	-9.798	< 0.001
DrvSenses	-0.0182	0.00262	3217	-6.936	< 0.001
BaseSenses	-0.0093	0.00157	3208	-5.916	< 0.001
DrvFreq*BaseFreq	0.0089	0.00102	3210	8.744	< 0.001
DrvFreq*DrvSenses	-0.0020	0.00136	3205	-1.457	0.145
BaseFreq*DrvSenses	-0.0004	0.00141	3208	-0.301	0.763
BaseFreq*BaseSenses	0.0006	0.00065	3196	0.906	0.365
DrvSenses*BaseSenses	0.0063	0.00090	3197	6.981	< 0.001
DrvFreq*BaseFreq*DrvSenses	0.0024	0.00061	3193	3.986	< 0.001
BaseFreq*DrvSenses*BaseSenses	-0.0018	0.00039	3189	-4.563	< 0.001

Table: Model output for analysis 2.

Analysis 3: Model

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	717.876	34.525	17.689	20.793	0.000
DrvFreq.c	-21.457	7.258	92.250	-2.956	0.004
logBaseResid.c	18.492	8.362	94.671	2.211	0.029
DrvNumSenses.c	-9.327	5.711	96.324	-1.633	0.106
BaseNumSenses.c	-4.119	2.606	96.918	-1.581	0.117
DrvFreq.c:DrvNumSenses.c	-6.548	2.496	92.214	-2.623	0.010
logBaseResid.c:DrvNumSenses.c	6.781	2.557	90.348	2.651	0.009
DrvNumSenses*BaseNumSenses	1.931	0.742	91.661	2.601	0.011