College of Arts & Sciences/Department of Linguistics

Do languages differ in semantic transparency of derived words? Using word vectors to explore English and Russian

Martha Booker Johnson, Andrea D. Sims, Micha Elsner

30

1 QUESTIONS

How does the frequency of a derived word and its base relate to semantic transparency? Is this relationship the same or different in English and Russian?

Our results complicate these issues, leading to a further question: Is it possible to disambiguate frequency and polysemy?

2 INTRODUCTION

Semantic Transparency

witness business smallness busyness

Low transparency — High transparency

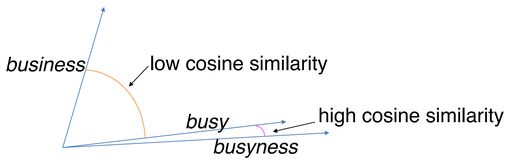
Existing Views on Semantic Transparency

"The belief that semantic transparency and the frequency of the derived form are linked is so widely held that it is sometimes stated as fact, without examples or references."[1]

"…in our experiments we consider relative frequency as a proxy of semantic transparency."[2]

3 WORD VECTORS

- Semantic transparency for thousands of derived/base pairs per language
- Distributional semantics via word vectors
  - Distributional semantics: word meanings are composed of the contexts where a word occurs[3]
  - Word vectors: based on words that occur around a word within a specified window[4]
- Higher cosine similarity = higher semantic transparency



3.1 METHODS

- List of derived/base pairs for each language
  - Eng: adapted from Celex[5]; 52 suffixes; 3372 pairs
  - Rus: adapted from Sims & Parker (2015)[6]; 19 suffixes; 3247 pairs
- Pre-trained lemmatized vectors from Fares et al. (2017)[7]
- Frequencies from Wikipedia & Russian National Corpus
- Cosine similarity for all pairs

3.2 RESULTS

Cosine similarity as a function of derived frequency

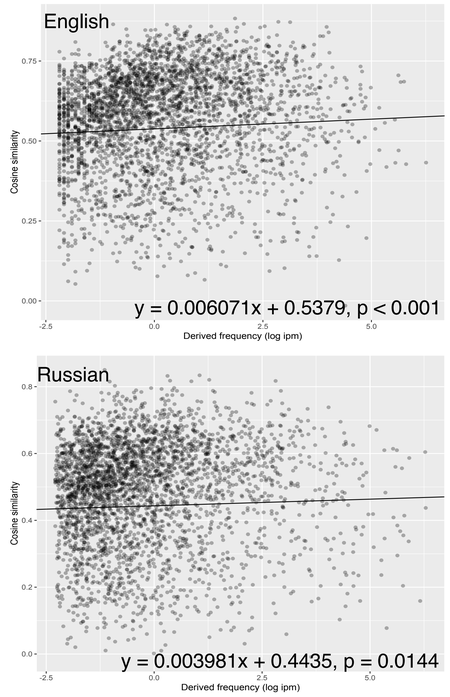


Figure 1. Both languages show a small, but significant, *positive* relationship between derived frequency and semantic transparency

- The claim that derived word has a *negative* relationship with semantic transparency does not immediately emerge in either language
- This prompts questions about what is meant by "semantic transparency"

4 EXPERIMENT COMPARISON

4.1 METHODS

- Compared cosine similarity to human semantic transparency judgments
- Judgments from McKenzie & Sims (2019)[8]
  - 24 English speakers, 25 Russian speakers
  - Prompt: How similar is the meaning of the word ***alienate*** to the meaning of the word ***alien***?
- Eng: 10 suffixes, 109 pairs; Rus: 9 suffixes, 91 pairs

4.2 RESULTS

Comparison of experimental & computational results

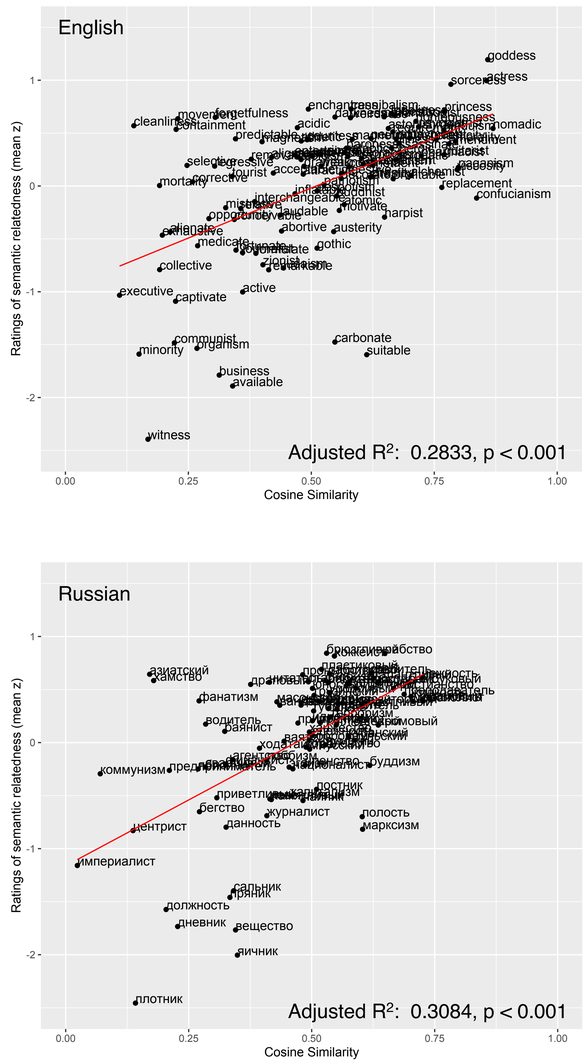


Figure 2. Cosine similarity and human semantic transparency judgements are correlated, but there is variability for low cosine similarity items

Polysemy of derived words as a function of frequency

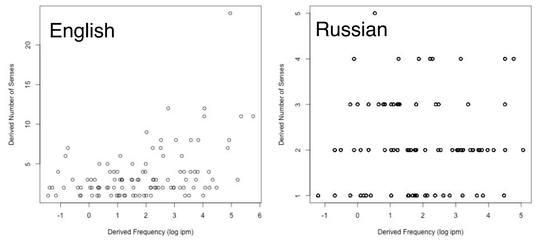


Figure 3. English has a significant positive relationship between frequency and polysemy, while a significant relationship does not exist in Russian

5 DISCUSSION

- Cosine similarity measurements amalgamate *all* meanings of a derived word relative to *all* meanings of its base. In contrast, our results suggest that human judgments identify semantically transparent meanings but ignore opaque relationships (Fig. 2).
- This suggests that polysemy complicates the relationship between frequency and transparency, which may explain why a negative relationship between frequency and transparency does not emerge (Fig. 1).
- Cross-linguistic comparisons of transparency need to consider differences in polysemy (Fig. 3).
- Exploring affix-specific trends[9] and deviation from the mean contrasts[2] will indicate whether some affixes are more subject to the above complications.

6 BIBLIOGRAPHY

1. Hay, J. (2001). Lexical frequency in morphology: Is everything relative? *Linguistics, 39*(6), 1041–1070. (p. 1043)
2. Varvara, R., Lapesa, G., & Padó, S. (2021). Grounding semantic transparency in context: A distributional semantic study on German event nominalizations. *Morphology*, 409–446. (p. 412)
3. Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2–3), 146–162.
4. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings.*
5. Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania.
6. Sims, A. D., & Parker, J. (2015). Lexical processing and affix ordering: cross-linguistic predictions. *Morphology*, *25*(2), 143–182.
7. Fares, M., Kutuzov, A., Oepen, S., & 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, 22-24 May 2017, Gothenburg, Sweden*, *131*, 271–276.
8. McKenzie, M., & Sims, A. D. (2019). Effects of frequency on word processing in Russian and English. *Paper Presented at the Fourth American International Morphology Meeting in Stony Brook, NY, May 3-5, 2019.*
9. Bonami, O., & Paperno, D. (2018). Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio*, *2*, 173–196.

THE OHIO STATE UNIVERSITY

ACKNOWLEDGMENTS

Thank you to Willy Cheung for sharing web-crawler code & Stephanie Antetomaso for technical help
Funding for conference travel provided by OSU Department of Linguistics' Student Travel Fund