

Explorations of morphological structure in distributional space


Harald Baayen,¹ Dunstan Brown,² and Yu-Ying Chuang¹

¹ University of Tübingen | ² University of York


This special issue brings together five studies that are the fruit of intense interactions between two research projects: The ‘Feast and Famine’ project funded by the UK’s Arts and Humanities Research Council, and the WIDE project funded by the European Research Council. The Feast and Famine project addresses overabundance and defectiveness in morphological paradigms. The WIDE project worked on a model of the mental lexicon and morphological processing in which form and meaning are represented by high-dimensional numeric vectors. What brought the two projects together is a shared interest in exploring the usefulness of distributional semantics for understanding morphology.

Distributional semantics, a research area at the intersection of artificial intelligence, psychology, and computational semantics, represents words’ meanings by means of high-dimensional vectors of real numbers calculated from large corpora. There are many ways in which such vectors, often referred to as ‘embeddings’, or ‘semantic vectors’, can be obtained. The latent semantic analysis (Landauer and Dumais, 1997) method first calculates how often words occur in documents, resulting in a word by document frequency table. Words that are similar in meaning or that are semantically related tend to occur in the same documents. As a consequence, the vector with a word’s document frequencies provides a semantic fingerprint of that word. As a second step, the word-document frequency table is subjected to a dimension reduction technique (singular value decomposition), resulting in a matrix of words by n latent dimensions. A typical value for n is 300. In short, LSA makes use of global statistics of how words co-occur across documents that cover a wide range of topics.

Various other methods use a sliding window technique that keeps track of the frequencies with which other words occur in the immediate context of a target word (e.g., HAL Burgess and Lund (1998); HiDEx, Shaoul and Westbury (2010); word2vec, Mikolov et al. (2013), and FastText, Bojanowski et al. (2017)). These methods build on the local statistics of words, rather than on their global statistics. FastText embeddings are available for a wide range of languages at <https://>

 Interactive figure available from <https://doi.org/10.1075/ml.00021.baa.figures>
<https://doi.org/10.1075/ml.00021.baa> | Published online: 12 September 2023

The Mental Lexicon ISSN 1871-1340 | E-ISSN 1871-1375

 Available under the CC BY 4.0 license. © 2023 John Benjamins Publishing Company

fasttext.cc/docs/en/crawl-vectors.html, and are considered an excellent choice for languages with rich morphological systems. The contribution on Finnish in the present special issue reports that indeed FastText outperforms word2vec for Finnish inflected nouns. Finally, GLoVe (Pennington et al., 2014) is a method for creating embeddings that leverages both global and local co-occurrence statistics. It is reported to outperform other methods on analogy tasks.

Embeddings have been found to be fruitful in the study of many aspects of lexical and morphological representation and processing. They have been used for predicting part-of-speech (Westbury and Hollis, 2018), basic emotions (Westbury et al., 2014), and personal relevance (Westbury and Wurm, 2022). Objective measures based on embeddings are now available for assessing semantic transparency, for example in compounding (Marelli et al., 2017; Shen and Baayen, 2021). Embeddings provide a fruitful starting point for quantitative modeling of conceptualization of a given meaning in terms of other meanings (Mitchell and Lapata, 2008), across inflectional morphology (Baayen et al., 2019), derivational morphology (Marelli and Baroni, 2015; Kisselew et al., 2015), and compounding (Marelli et al., 2017). Williams et al. (2019) used FastText embeddings from languages without gender to create a basis for investigating the degree of semantic arbitrariness in gender assignment in languages with gender. Guzmán Naranjo (2020) used vectors to model semantics in an analogical classification approach to Russian noun inflection. Bonami and Paperno (2018) used embeddings to clarify differences between inflectional and derivational morphology. Baayen and Moscoso del Prado Martín (2005) used embeddings to document differences in the meanings of English regular and irregular verbs, and Heitmeier and Baayen (2020) used them to model the problems that arise in aphasia with regular and irregular verbs. Within the framework of the discriminative lexicon model (Baayen et al., 2019), a mapping from speech audio to embeddings can be trained on existing words, and used to predict the meanings of auditory nonwords. The embeddings predicted for these nonwords are in turn informative about both reaction times to these nonwords in auditory lexical decision and the spoken word durations of these nonwords (Chuang et al., 2020). Corpus-based embeddings were used by Nieder et al. (2022) to model noun plurals in Maltese, and by Chuang et al. (2022) to model morphological priming (see also Marelli et al., 2013). To probe the relation between form and meaning, Marelli et al. (2014) and Amenta et al. (2019) proposed form-to-meaning consistency measures that build on embeddings and shed new light on many priming experiments.

The studies on morphology that are brought together in the present special issue build on and are inspired by the experiences with embeddings that have accumulated in the literature. Before providing an overview of the contributions

of the individual studies, we provide some technical background on working with embeddings.

Assessing vector similarity

Vectorial representations of word meanings have the advantage that they quantify degrees of semantic similarity. The similarity between vectors can be assessed in many ways. Consider Figure 1, which represents four sets of English nouns in a 3-dimensional space. A dynamic version of this figure is available at: <https://doi.org/10.1075/ml.00021.baa.video1>, and a more detailed interactive figure can be found at: <https://doi.org/10.1075/ml.00021.baa.fig1>. In this figure, the points can be rotated using the left mouse button, and hovering with the mouse above a data point will bring up the word and its semantic class. All contributions to this special issue provide links to interactive graphics, and the reader is encouraged to follow these links, as the interactive versions of figures are much more informative.

Returning to Figure 1, words for plants are found in the upper back corner, and are presented in orange. Words for body parts are located near the lower right corner, and are shown in purple. Words for people (professions) are presented in green, and are close to the center of the bottom plane. Words for animals (in purple) cluster in the lower center back. These clusters (which emerge from corpus-based vectors using principal components analysis) illustrate that words that are similar in meaning will tend to occur close together in distributional space.

Proximity in distributional space can be measured in many ways. One possibility is to consider the distance between two points. For instance, *scientist* and *coconut* are far away from each other, and we can use the Euclidean distance to measure the length of the vector starting at *scientist* and ending at *coconut*. Given the vector x for *scientist* and the vector y for *coconut*, the Euclidean distance $d(x, y)$ between these two vectors is given by

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}, \quad (1)$$

where x_i and y_i are the components of x and y . In the three-dimensional space of Figure 1, the number of dimensions n is equal to 3, but in the original space, $n=300$, and it is the distance in this much higher dimensional space that is usually of interest. Several contributions to this special issue make use of the Euclidean distance measure in order to get a sense of where in distributional space words are located, and how they cluster.

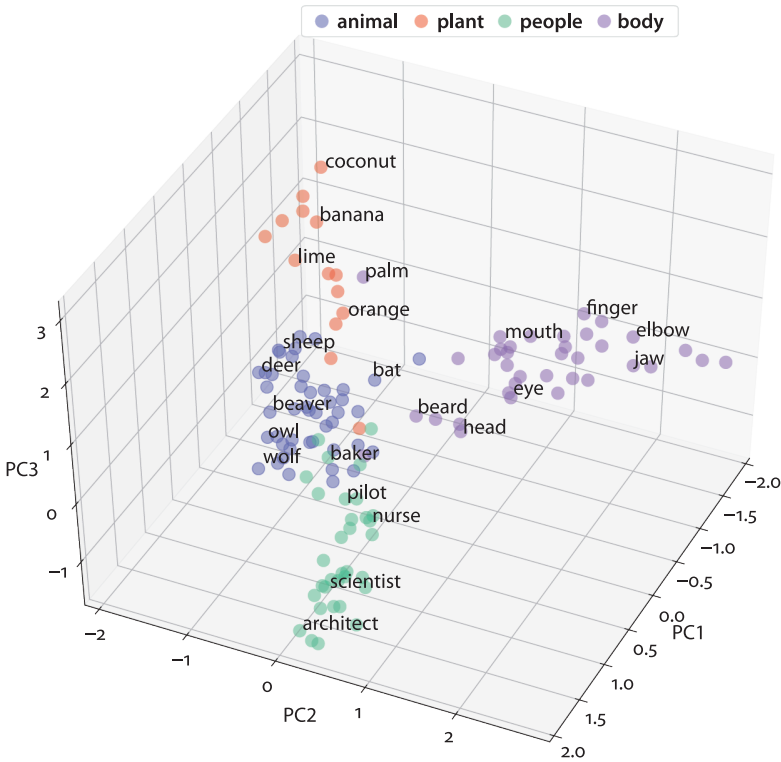


Figure 1. A principal components analysis of a selection of 300-dimensional wordvec embeddings of English nouns reveals clustering by semantic category. More similar words are found closer together. Since standard embeddings for homographs are identical, their position in distributional space can be suboptimal. In this example, *palm*, although color-coded as a body part, clusters with the plants and fruits, suggesting that this word is predominantly used in texts as a tree or fruit

Two related measures, cosine similarity and Pearson correlation, tend to correlate better with human intuitions of semantic similarity than Euclidean distance does. Since both measures are used in the studies brought together in this special issue, we briefly introduce both.

Cosine similarity is illustrated for *scientist* and *coconut* in Figure 2, using a 2-D plane for ease of presentation. The angle between the vector x to *scientist* and the vector y to *coconut* is wide and close to 90 degrees in this 2-D projection. The cosine of this angle is therefore close to zero. Conversely, the angle between x and the vector z to *nurse* is close to zero degrees, and hence the cosine of this angle is close to 1. The cosine of the angle between two vectors thus captures the degree to which two vectors point in the same direction. For an n -dimensional space, the cosine similarity $S_C(x, y)$ is defined as:

$$S_C(x, y) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}} \tag{2}$$

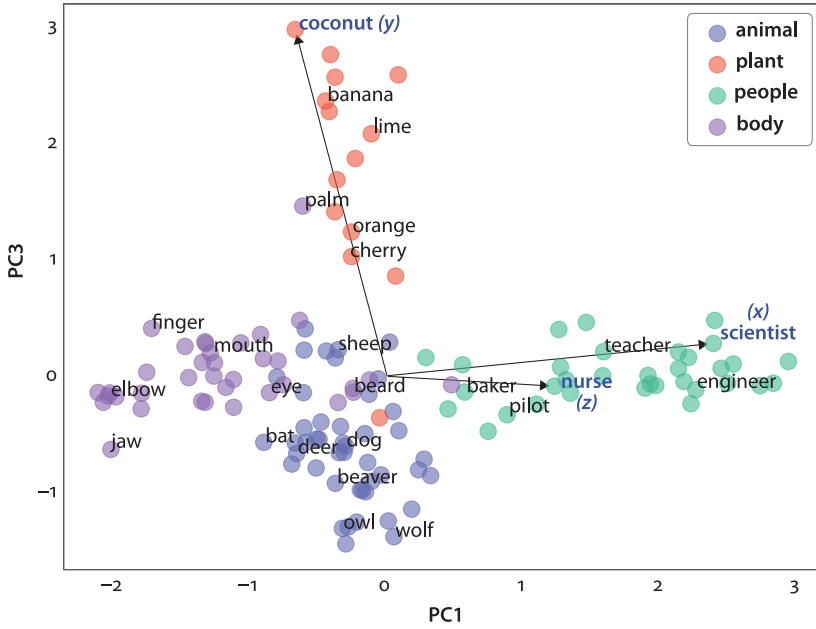


Figure 2. The angle between the vectors for *scientist* and *coconut* is large, and the cosine of the angle is small, indicating lack of similarity. Conversely, the angle between the vectors of *scientist* and *nurse* is small, and the cosine of the angle is large, indicating greater similarity

When the vectors x and y are centered by subtracting their means (\bar{x} and \bar{y} respectively), we obtain the Pearson correlation.

$$r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \tag{3}$$

In other words, the correlation is the centered cosine similarity. If two vectors have more similar coordinates, and hence coordinates that are more strongly correlated, they will point to more similar locations in distributional space. In practice, the cosine similarity and the Pearson correlation produce slightly different, but highly similar, values.

Dimension reduction, visualization, and classification

All the studies in this special issue wrestle with the question of how to make sense of word embeddings. To address this question, they make heavy use of visualization. As mentioned above, many of the figures have links to dynamic visualizations, and readers are encouraged to explore these interactive plots in order to get a better sense of the structure of the embedding spaces.

Visualization of high-dimensional embedding spaces requires some form of pre-processing so that data-points can be presented in a two-dimensional plane or a three-dimensional cube. The two figures above made use of principal components analysis, a technique that sets up new axes (principal components) that are ordered by the amount of variance they capture. Sometimes, these dimensions are interpretable. In Figure 2, the first principal component (PC₁) pulls apart body parts (left) and people (right), with animals and plants positioned roughly where PC₁ is zero.

A technique used across all the studies reported here is t-distributed stochastic neighborhood embedding (t-SNE, Maaten and Hinton, 2008), which is an unsupervised clustering method with a high likelihood of visually representing clusters in 2D or 3D maps if clusters are truly present in the high-dimensional space. All that the t-SNE algorithm does is provide new coordinates for multi-dimensional data points in a low-dimensional space. In order to see whether clusters exist, the analyst will typically use color coding for theoretically motivated classes in 2D or 3D scatterplots, as illustrated in the figures above. The studies brought together in this special issue contribute to a better understanding of what t-SNE does, but as t-SNE has not yet enjoyed wide usage in linguistic research and research on the mental lexicon, all analyses remain tentative and exploratory.

When the interest is in a low-dimensional space that (unlike t-SNE) respects, as much as possible, the distances between points in the original space, classical multi-dimensional scaling is an excellent choice.

It can be useful to complement unsupervised clustering with supervised classification. A classifier that is used in one of the contributions to this special issue is linear discriminant analysis. Although this method is roughly 80 years old, it works surprisingly well for predicting morphological properties of word embeddings.

Distributional vectors: What are they actually?

Various authors contributing to this special issue have mixed feelings about embeddings. What do word embeddings actually represent? What do the dimen-

sions of embeddings represent? Are embeddings sophisticated enough to capture linguistically relevant generalizations?

In artificial intelligence and natural language processing, embeddings have become an indispensable workhorse. Clearly, representing words' "meanings" as high-dimensional numeric vectors informed by how words are used in text has significant advantages. At the same time, the volumes of text from which embeddings are calculated far surpass the experience of individual human language users. This suggests that embeddings provide representations that approximate human knowledge, but do so in ways that do not necessarily reflect human learning and cognition.

Whereas some researchers regard embeddings as tools that are better than nothing, but that cannot be really trusted, others suggest that they capture the essence of meaning. For instance, Westbury and Hollis (2018) argue, using embeddings, that part-of-speech categories are actually semantic categories. Others, however, will argue that embeddings capture a mixture of lexical semantics, syntax, pragmatics, registers, and styles, and that which parts of this mixture dominate in "semantic" vectors can vary markedly from language to language. The contributions to this special issue use the terms 'embeddings' and 'semantic vectors' interchangeably, and understand these terms as denoting knowledge of words' use that is independent of words' forms. All the contributions, to which we now turn, have found that embeddings enrich our insights into the relationship between semantics, syntax, and morphology.

Overview of contributions

The first study in this special issue investigates English noun plurals. Its starting point is the observation that the change in distributional space from a singular to its plural depends on semantic class. For instance, plurals of animals shift away from their singulars in a different direction than the plurals of plants. This raises the question of how to best model the conceptualization of noun plurals in English: Is a linear transformation from singulars to plurals sufficiently powerful to model the observed dependency on semantic class, or do we need to explicitly condition on semantic class? These questions are addressed by an investigation of how well conceptualized plural vectors align with their spoken words, using a large multimodal corpus of conversational English.

The second study addresses the inflection of Finnish nouns. Finnish nouns are inflected for 14 cases and 2 numbers. In addition, a possessive suffix can be attached, as well as various clitics. Inspection of Finnish inflected nouns with FastText embeddings, using t-SNE as unsupervised clustering technique, revealed

clustering by case, and within case, clustering by number. Possessive suffixes and clitics showed some additional clustering. The authors decomposed empirical vectors for Finnish nouns into imputed vectors for lexemes, cases, numbers, possession, and the discourse functions realized by Finnish clitics. A comparison of a series of decompositional models revealed that decompositions are more precise when they allow for interactions between inflectional features, such as the interaction of plurality by case. This result raises the question of how these interactions are best accounted for in theories of morphology.

The third study addresses noun paradigms in Russian, which inflect for 6 cases and 2 numbers. As for Finnish, unsupervised clustering with t-SNE reveals grouping first by case, and within case, by number. In other words, the first three studies of this special issue all provide evidence of plural semantics being realized conditionally, on semantic class in English, and on case in Finnish and Russian. The central topic of this third study is the defectivity of the paradigms of a small subset of Russian nouns, which lack a genitive plural. It appears that the distributional vectors of the inflected variants of these defective nouns are more idiosyncratic than those of other nouns, suggesting that possibly there is more to defectivity than just uncertainty associated with the realization in form.

The last two studies move from inflectional morphology to word formation, studying derivational morphology in German and compounding in Mandarin. The fourth study documents two remarkable properties of German particle verbs. First, their category-conditioned productivity (P) decreases linearly with their extent of use (profitability) (V), suggesting that all particle verbs are sampled from basically the same population. Second, particle verbs do not form clusters by particle in t-SNE maps. In contrast, most German affixed words do cluster by affix in t-SNE maps, and show no correlation between P and V , indicating that affixed words with different affixes come from very different populations. These results suggest that German particles may be rather undifferentiated in their semantics, basically providing different forms for realizing very similar meanings. However, an analysis with supervised classification shows that there *is* enough information in the embeddings to accurately predict particles; upon closer inspection, this information turns out to be spread out across all dimensions, which apparently makes it invisible to the t-SNE analysis.

The final study investigates the productivity and transparency of compounding in Mandarin Chinese, and compares compounding with suffixation. Mandarin compounds reveal quantitative properties similar to those of German particle verbs, and Mandarin suffixed words have properties that mirror those of German affixed words. An analysis of the geometry of compounding and suffixation suggests a potential measure that predicts both semantic transparency and the probability of new forms, namely the extent to which clusters of words

sharing the same constituent are homogeneous, in the sense that these clusters are not invaded by intruders from other constituent clusters. In fact, constituents with very few intruders from other constituents show up with some clustering in t-SNE maps, and resemble suffixes (which also have low intruder ratios). This leads to the conclusion that Mandarin compounds realize, at one extreme, motivated but semantically unsystematic concept formation (where other constituents could just as well have been used), and at the other extreme, systematic suffix-like semantics.





Concluding remarks

This special issue has profited immensely from the feedback received from the reviewers and the editors, Melanie Bell, Juhani Järvikivi, and Vito Pirrelli. Their criticism and feedback has helped us as writers to clarify our thinking, to better understand our findings, to think through alternative interpretations, and to communicate our findings more clearly. The help we received from our reviewers and editors is deeply appreciated by all authors.

References

- Amenta, S., Crepaldi, D., and Marelli, M. (2019). Consistency measures individuate dissociating semantic modulations in priming paradigms: A new look on semantics in the processing of (complex) words. *Quarterly Journal of Experimental Psychology*, 73(10), 1546–1563.
- Baayen, R.H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*.
- Baayen, R.H. and Moscoso del Prado Martín, F. (2005). Semantic density and past-tense formation in three Germanic languages. *Language*, 81:666–698.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Bonami, O. and Paperno, D. (2018). Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio*, 17(2):173–195.
- Burgess, C. and Lund, K. (1998). The dynamics of meaning in memory. In Dietrich, E. and Markman, A.B. editors, *Cognitive dynamics: Conceptual change in humans and machines*. Lawrence Erlbaum Associates.
- Chuang, Y., Vollmer, M.-L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., and Baayen, R.H. (2020). The processing of pseudoword form and meaning in production and comprehension: A computational modeling approach using linear discriminative learning. *Behavior Research Methods*. 53: 945–976.

- [doi](#) Chuang, Y.Y., Kang, M., Luo, X.F., and Baayen, R. H. (2023). Vector space morphology with linear discriminative learning. In Crepaldi, D. editor, *Linguistic morphology in the mind and brain*. Routledge.
- [doi](#) Guzmán Naranjo, M. (2020). Analogy, complexity and predictability in the Russian nominal inflection system. *Morphology*, 30(3):219–262.
- [doi](#) Heitmeier, M. and Baayen, R. H. (2020). Simulating phonological and semantic impairment of English tense inflection with Linear Discriminative Learning. *The Mental Lexicon*, accepted. *PsyArXiv*. 15(3): 385–421.
- Kisselew, M., Pado, S., Palmer, A., and Šnajder, J. (2015). Obtaining a better understanding of distributional models of german derivational morphology. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 58–63.
- [doi](#) Landauer, T. and Dumais, S. (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.
- Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605.
- [doi](#) Marelli, M., Amenta, S., and Crepaldi, D. (2014). Semantic transparency in free stems: the effect of orthography-semantics consistency in word recognition. *Quarterly Journal of Experimental Psychology*, 68(8): 1571–1583.
- [doi](#) Marelli, M., Amenta, S., Morone, E.A., and Crepaldi, D. (2013). Meaning is in the beholder’s eye: Morpho-semantic effects in masked priming. *Psychonomic bulletin & review*, 20(3):534–541.
- [doi](#) Marelli, M. and Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, 122(3):485–515.
- [doi](#) Marelli, M., Gagné, C.L., and Spalding, T.L. (2017). Compounding as abstract operation in semantic space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition*, 166:207–224.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 26.
- Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *proceedings of ACL-08: HLT* (pp. 236–244).
- [doi](#) Nieder, J., Chuang, Y.Y., van de Vijver, R., & Baayen, H. (2023). A discriminative lexicon approach to word comprehension, production, and processing: Maltese plurals. *Language*, 99(2), 242–274.
- [doi](#) Pennington, J., Socher, R., and Manning, C.D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.
- [doi](#) Shaoul, C. and Westbury, C. (2010). Exploring lexical co-occurrence space using hidex. *Behavior Research Methods*, 42(2):393–413.
- [doi](#) Shen, T. and Baayen, R. H. (2021). Adjective-noun compounds in Mandarin: a study on productivity. *Corpus Linguistics and Linguistic Theory*. 18(3), 543–572.

-  Westbury, C. and Hollis, G. (2018). Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behavioral Research Methods*, 51: 1371–1398.
-  Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., and Jacobs, A. M. (2014). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *The Quarterly Journal of Experimental Psychology* 68: 1599–1622.
-  Westbury, C. and Wurm, L. H. (2022). Is it you you're looking for? personal relevance as a principal component of semantics. *The Mental Lexicon*, 17(1):1–33.
-  Williams, A., Blasi, D., Wolf-Sonkin, L., Wallach, H., and Cotterell, R. (2019). Quantifying the Semantic Core of Gender Systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5734–5739, Hong Kong, China. Association for Computational Linguistics.

Address for correspondence

Harald Baayen
University of Tübingen
harald.baayen@uni-tuebingen.de

Co-author information

Dunstan Brown
University of York
dunstan.brown@york.ac.uk

Yu-Ying Chuang
University of Tübingen
yu-ying.chuang@uni-tuebingen.de

Publication history

Date received: 1 December 2022
Date accepted: 23 December 2022
Published online: 12 September 2023