

This is a repository copy of *Explorations of morphological structure in distributional space*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/203772/>

Version: Published Version

---

## Article:

Baayen, Harald, Brown, Dunstan [orcid.org/0000-0002-8428-7592](https://orcid.org/0000-0002-8428-7592) and Chuang, Yu-Ying (2023) Explorations of morphological structure in distributional space. *The Mental Lexicon*. ISSN 1871-1340

<https://doi.org/10.1075/ml.00021.baa>

---

## Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

## Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing [eprints@whiterose.ac.uk](mailto:eprints@whiterose.ac.uk) including the URL of the record and the reason for the withdrawal request.

# Explorations of morphological structure in distributional space


Harald Baayen,<sup>1</sup> Dunstan Brown,<sup>2</sup> and Yu-Ying Chuang<sup>1</sup>

<sup>1</sup> University of Tübingen | <sup>2</sup> 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](https://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

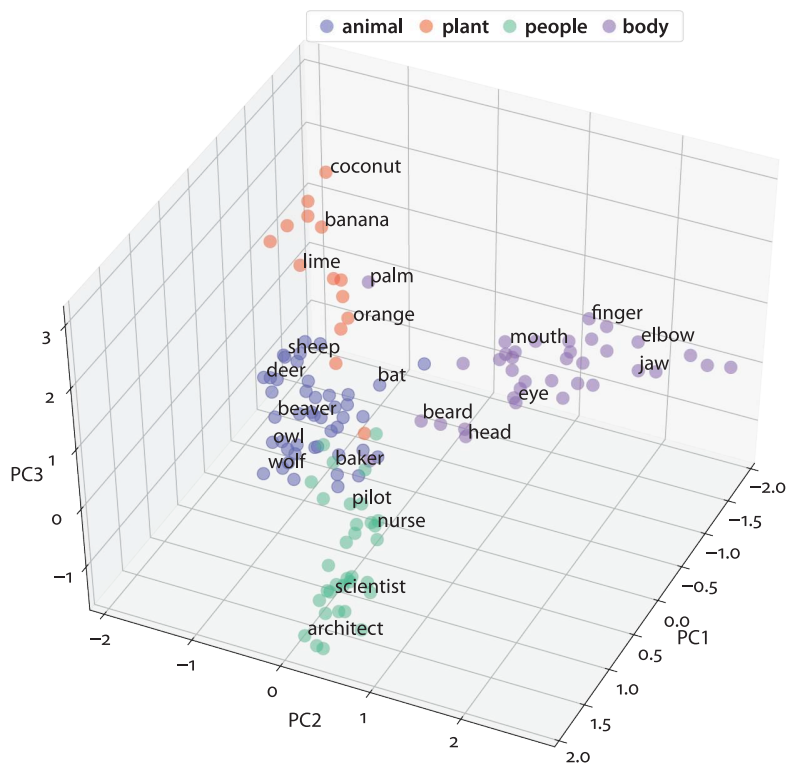
Vectorial representations of word meanings have the advantage that they quantify degrees of semantic similarity. The similarity between vectors can be assessed in many ways. Consider Figure 1, which represents four sets of English nouns in a 3-dimensional space. A dynamic version of this figure is available 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 word2vec 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:

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

(2)

