

## case 23 dezubicaray senses

Participant-visible method context. Results, discussion, captions, and the target figure are excluded.

**Materials** All English words with at least one sense listed in the WordNet database (monosemes and polysemes; Gao et al., 2023; Miller, 1995) were selected for investigation. The subset of lemmas was first derived using the `textstem` and `lexicon` packages in R Version 4.3.1 (R Core Team, 2023; Rinker, 2018a, 2018b) excluding all polymorphemic forms according to the MorphoLex database (SánchezGutiérrez et al., 2018) to prevent morpho-phonological redundancies influencing the analyses (e.g., affixation). Next, 969 homonyms were excluded according to the lists compiled by Rice, Beekhuizen et al. (2019) and Hauer and Kondrak (2020), as well as an additional 338 homophones (words with different spelling and distinct meanings but identical pronunciation, e.g., hair-hare, flour-flower) using the `LexFindR` package (Li et al., 2022). Words with shared etymological roots were not excluded given the aim was to investigate relationships between phonological features and polysemy (number of senses per word) rather than between different word forms with similar meanings. In addition, as Dautriche, Mahowald, Gibson, and Piantadosi (2017) noted, many words with shared historical roots now have divergent meanings (e.g., “shadow” and “shed” are from the same Old English root). Finally, forms that comprised more than three syllables or were numbers, proper names, function words, prepositions, and interjections according to Brysbaert et al.’s (2012) dominant part of speech classification were also excluded. Forms with four or more syllables were excluded as they comprised only 159 monomorphemic words and to facilitate comparison with Sharpe and Marantz’s (2017) measure of noun/verb typicality that was based on three or fewer syllables. Phonemic transcriptions were retrieved from the Carnegie Mellon University pronouncing dictionary (<https://www.Speech.cs.cmu.edu/cgi-bin/cmudict>) and stress category assignments from Tucker et al.’s (2019) database based on the same. Each monomorphemic word was coded according to its whole word properties (number of letters, syllables, and phonemes), their initial and end phonemes, number of phonetic features and their initial and final positions (i.e., place and height for vowels; voicing; place and manner of articulation for consonants), and syllabic stress position (initial, medial, final, multiple). This process resulted in a total of 4,466 monomorphemic words (839 monosemes) for each of which 60 phonological/phonetic variables were calculated, with the addition of orthographic length (i.e., number of letters).<sup>3</sup> Note that phonemes were not additionally coded into subsyllabic onset, nucleus, and coda slots as has been the approach used by some (e.g., Monaghan et al., 2014) but not other studies investigating the phonological similarities of word pairs (e.g., Dautriche, Mahowald, Gibson, & Piantadosi, 2017; Shillcock et al., 2001; Tamariz, 2008). Doing so with English polysyllabic words requires strong theoretical assumptions about where syllable boundaries occur, such as the “maximal onset principle” that assigns consonants between vowels to the onsets of the following vowel rather than the coda, despite the fact that linguists have argued there is no real evidence to support the

maximal onset principle as many English words do not conform to this structure (e.g., Duanmu, 2009). Further, as certain phonemes/ phoneme clusters tend to co-occur in English due to phonotactic constraints, this approach would also introduce collinearity into the variable set.

**Design and Analysis** All analyses described below were performed in R Version 4.3.1 (R Core Team, 2023). A multistep approach was adopted to predict polysemy with form variables (de Zubicaray et al., 2023). First, the WordNet number of senses variable was log transformed due to its skewed distribution (see also Hoffman et al., 2013), and any form variables with linear dependencies were excluded using the `caret` package (Kuhn, 2022). Next, the `leaps` package (Lumley, 2022) was used to determine the best subsets of form variables for predicting polysemy. A 10-fold cross-validation (repeated 200 times with different randomized folds) approach was then adopted using the `caret` package with 90% of the data as training data and 10% as test, selecting the best-fit model in terms of predictive accuracy and model complexity based on its minimal root-mean-square error. This approach to model selection is preferable as it prevents overfitting (see de Rooij & Weeda, 2020; Yarkoni & Westfall, 2017). In a final step, the best-fit model was entered into a linear regression with robust standard errors (Wilcox, 2019) using the `estimatr` package (Blair et al., 2022), to ensure valid coefficient estimates in the presence of any multicollinearity, heteroscedasticity, skewness, or outliers. Next, to establish that the systematic polysemy-form mappings identified by the best-fit model were not spurious, the best subsets and cross-validation steps were repeated after randomly shuffling the log WordNet number of senses among the form variables. To derive a measure of polysemy-form mapping, the predicted value of the dependent variable for each word according to the best-fit model was extracted (i.e., the fitted values from the matrix of predicted means). These values were

Z-transformed so that forms typical of multiple senses are indicated by positive values and those with fewer senses (tending to monosemy) by negative values.

Transparency and Openness All data and analysis scripts are available at <https://osf.io/t4zyj> (de Zubicaray, 2025). The study was not preregistered.