

Sensorimotor distance: A grounded measure of semantic similarity for 800 million concept
pairs

Cai Wingfield¹ & Louise Connell^{1, 2}

¹ Department of Psychology, Lancaster University

² Department of Psychology, Maynooth University

Author Note

Cai Wingfield <https://orcid.org/0000-0002-0254-199X>

Louise Connell <https://orcid.org/0000-0002-5291-5267>

Correspondence concerning this article should be addressed to Cai Wingfield,
Department of Psychology, Fylde College, Lancaster University, Lancaster LA1 4YF, UK or
to Louise Connell, Department of Psychology, Maynooth University, Maynooth, Co. Kildare,
Ireland. Email c.wingfield@lancaster.ac.uk or louise.connell@mu.ie

Abstract

Experimental design and computational modelling across the cognitive sciences often rely on measures of semantic similarity between concepts. Traditional measures of semantic similarity are typically derived from distance in taxonomic databases (e.g. WordNet), databases of participant-produced semantic features, or corpus-derived linguistic distributional similarity (e.g. CBOW), all of which are theoretically problematic in their lack of grounding in sensorimotor experience. We present a new measure of *sensorimotor distance* between concepts, based on multidimensional comparisons of their experiential strength across 11 perceptual and action-effector dimensions in the Lancaster Sensorimotor Norms. We demonstrate that, in modelling human similarity judgements, sensorimotor distance has comparable explanatory power to other measures of semantic similarity, explains variance in human judgements which is missed by other measures, and does so with the advantages of remaining both grounded and computationally efficient. Moreover, sensorimotor distance is equally effective for both concrete and abstract concepts. We further introduce a web-based tool (<https://lancaster.ac.uk/psychology/smdistance>) for easily calculating and visualising sensorimotor distance between words, featuring coverage of nearly 800 million word-pairs. Supplementary materials are available at <https://osf.io/d42q6/>.

Keywords: sensorimotor distance; semantic similarity; semantic distance; grounded cognition

Sensorimotor distance: A grounded measure of semantic similarity for 800 million concept pairs

Semantic similarity is at the heart of many fundamental processes in human cognition (see Goldstone & Son, 2012; Hahn, 2014), such as categorisation (e.g., Hampton, 1998; Nosofsky, 1986), memory recall and recognition (e.g., Baddeley, 1966; Montefinese et al., 2015) and language processing (e.g., Raveh, 2002; Hutchison et al., 2008). As semantic similarity between concepts is known to have such wide-ranging effects, accurate and interpretable measures of similarity are crucial tools for predicting behaviour and mapping out how people conceptualize and process their world. Thus, such measures are of the utmost utility in research in the cognitive sciences, from designing and analysing experiments to building computational models.

Study of semantic similarity is inseparable from theories of conceptual processing and representation in general. For theories that assume the conceptual system is organised in a taxonomic hierarchy (e.g. Collins & Quillian, 1969; Jolicoeur et al., 1984), natural candidates for measures of semantic similarity may be derived from measures of distance in hierarchical databases (e.g. WordNet, Princeton University, 2010). On the other hand, for family-resemblance accounts in which conceptual relationships are founded on shared features (Rosch & Mervis, 1975; Wittgenstein, 1953; see also Cree & McRae, 2003), measures of semantic similarity can be produced by comparing lists of conceptual features produced in norming studies (e.g. Buchanan et al., 2019; McRae et al., 2005). Under the distributional hypothesis—that similarity of word meaning is given by similarity of word usage (Harris, 1954; Firth, 1957)—linguistic distributional measures of semantic similarity may be derived by extracting relevant statistics from large corpora of natural language (e.g. latent semantic analysis, LSA: Landauer & Dumais, 1997; continuous bag of words, CBOW: Mikolov et al.,

2013). Finally, within a grounded cognition framework, where concepts' representations involve partial replay of perception and action experience (e.g. Barsalou, 1999; Connell & Lynott, 2014), one might surmise that similarity between concepts equates to similarity of their sensorimotor experience; however, no measure of semantic similarity based on sensorimotor experience has been made available to date.

Our goal in the present paper is to address that gap by providing a database of semantic similarity measures based on the sensorimotor experience underlying each concept, which we term *sensorimotor distance* (available online at <https://lancaster.ac.uk/psychology/smdistance>).

Measures of Semantic Similarity

While different measures of semantic similarity have tended to emerge from different (and often conflicting) theoretical traditions, it does not mean they are mutually exclusive. Similarity is a multifaceted and complex construct. For instance, if two things are similar because they share properties in common, then similarity itself is meaningless because all objects share an infinite number of properties in common (e.g., a *plum* and a *lawnmower* both share the properties of weighing less than 100 kg, and less than 101 kg, and less than 102 kg, etc.: Goodman, 1972). Similarity is thus only meaningful when it is constrained to mean two things are similar *in a certain respect*, and it is possible that multiple measures of similarity, each applying different constraints, are required to fully capture the similarity between two given concepts.

Similarity measures based on hierarchical structure can be taken from large machine-searchable encyclopaedic databases (e.g., Strube & Ponzetto, 2006), or purpose-built semantic databases such as WordNet (Princeton University, 2010; Miller, 1995, 1998). WordNet is a large online lexical database of English, with words organised into a hierarchy of hypernymic (i.e., "is a type of") relations. Under this framework, concepts are more

similar when there is a short path between their nodes in the hierarchical structure (Resnik, 1995; Jiang & Conrath, 1997). For example, similarity measures based on WordNet distance are likely to score *alligator* and *crocodile* as highly similar because the path between them is very short (e.g., *alligator* → *crocodilian reptile* → *crocodile*), but will score *alligator* and *monster* as quite dissimilar because the path between them requires going via the root node of *entity* and is thus very long indeed¹. Coverage of similarity comparisons using WordNet distance is very high in principle (i.e., over 117,000 synset classes potentially enables billions of pairwise comparisons), although it is limited to separate consideration of nouns and verbs because other parts of speech are not structured in hypernymic hierarchies; “off the shelf” coverage is far smaller in reality, such as Maki et al.’s (2004) compilation of WordNet distances for nearly 50,000 concept pairs. However, although the nature of hierarchical distance as a similarity measure means that while it excels at constraining similarity by hypernymic/categorical relations, the role of sensorimotor grounding is largely non-existent. While concepts very close together in hierarchical structure may share some sensorimotor experience (e.g. many types of *foodstuff* may be grounded in taste and smell), other forms of semantic similarity that are grounded in perceptual or action resemblances (e.g. *alligator* and *monster*; *princess* and *bride*; *toddler* and *detonation*) are not generally captured.

Feature-based similarity measures, on the other hand, are typically computed from lists of features produced by participants per concept in norming studies. Under this framework, similarity between a pair of concepts is given by the degree of overlap of their respective lists of features. Feature lists are necessarily highly sparse (i.e., most concepts do not possess most features); overlap can therefore be determined by simple counting of common features (McRae et al., 2005), or incorporating feature-production frequencies (e.g.,

¹ e.g., *alligator* → *crocodilian reptile* → *diapsid reptile* → *reptile* → *vertebrate* → *chordate* → *animal* → *organism* → *living thing* → *whole* → *physical object* → *physical entity* → *entity* → *abstract entity* → *psychological feature* → *cognition* → *ability* → *creativity* → *imagination* → *imaginary being* → *monster*.

by using the cosine of the angle between feature-frequency vectors: Devereux et al., 2014; Buchanan et al., 2019). For example, the concepts *mountain* and *hill* would be scored as similar because they have many shared features such as *high*, *landscape*, *climb* and *steep*, whereas *mountain* and *pyramid* would be far less similar because they share far fewer features (e.g., *tall*). By encompassing a wide range of features including taxonomic (e.g. *landscape*), encyclopaedic (e.g. *found in ranges*) and grounded (e.g. *cold*), feature-based measures can theoretically constrain similarity on a number of different dimensions. However, grounded features are not consistently present across concepts (e.g., *toy* has no perceptual or action features in McRae et al.'s norms; *music* has no action features in Buchanan et al.'s norms), and so a measure of semantic similarity based on concept-feature norms is, at best, inconsistently and partially grounded. In addition, the laborious nature of collecting and standardising feature lists produced by participants has meant that feature-based similarity measures are quite restricted in their coverage. One of the largest concept-feature norming studies is that of Buchanan et al. (2019), who compiled a database of features for almost 4,500 concepts that expanded on several previous databases (including Devereaux et al., 2014; McRae et al., 2005; Vinson & Vigliocco, 2008), and made available feature-based similarity measures for over 200,000 concept pairs. While useful, feature-based measures nonetheless cover only a small fraction of the approximately 40,000 concepts thought to make up the typical conceptual system of adult English speakers (Lynott et al., 2020; see also Brysbaert et al., 2014) and a smaller fraction of the hundreds of millions of comparable concept pairs. Moreover, since many concept-feature norming studies focused exclusively on concrete noun concepts, particularly objects (e.g., Devereaux et al., 2014; McRae et al., 2005), and later studies expanded those item sets (Buchanan et al., 2019), abstract concepts and other parts of speech remain underrepresented in feature-based similarity measures.

Finally, linguistic distributional measures of semantic similarity are based on the statistical relationships between words and their usage contexts in natural language. Under this framework, similarity of concepts is determined by contextual similarity of their word labels, following the distributional hypothesis that words with similar meanings tend to occur in similar contexts (Harris, 1954). Linguistic distributional measures of semantic similarity are recently typified by Mikolov et al.'s (2013) continuous bag of words (CBOW), which represents words as vectors derived from a neural network model trained on word co-occurrences in a corpus of text; similarity between two concepts is then compared as the cosine similarity between these vectors. For example, CBOW scores *helicopter* and *airplane* as highly similar because they appear in similar contexts (e.g., concerning *pilot*, *flying*, *sky*), but scores *helicopter* and *bee* as dissimilar because they tend to occur in quite different contexts. Other examples of linguistic distributional measures include latent semantic analysis (LSA: Landauer & Dumais, 1997, which continues to be used extensively in the cognitive sciences as a measure of semantic similarity), GloVE (Pennington et al., 2014), and skip-gram (Mikolov et al., 2013: CBOW's sister model in the word2vec package). Linguistic distributional measures of semantic similarity have excellent coverage, with tens or hundreds of thousands of individual words available for comparison (depending on the corpus) that span all parts of speech. They also appear to constrain similarity on a number of different dimensions, such as synonymy, shared categories, taxonomic classes, and thematic connections (see Wingfield & Connell, 2022, for review). However, linguistic distributional measures can approximate sensorimotor grounding only insofar as this information is reflected in statistical patterns of word usage, which is limited. For example, Louwerse & Connell (2011) showed that language-use statistics were able to distinguish visuohaptic words from auditory words, but not visual words from haptic (see also Louwerse & Jeuniaux, 2008; Riordan & Jones, 2011). In general, linguistic distributional measures do not capture

many forms of semantic similarity that are grounded in perceptual or action resemblances² (e.g. *helicopter and bee, toddler and detonation*).

The Current Norms: Sensorimotor Distance

We present here a novel, grounded measure of semantic similarity: *sensorimotor distance*. It is based on the Lancaster Sensorimotor Norms (Lynott et al., 2020), which contain sensorimotor strength ratings that reflect the extent to which a given referent concept can be perceived through auditory, gustatory, haptic, interoceptive, olfactory, and visual modalities; or can be experienced by performing an action with the hand/arm, head, foot/leg, mouth, or torso effectors. Each of these dimensions was carefully chosen to map to a specific, separable region of the cortex, meaning that a multidimensional profile of sensorimotor strength approximates the distributed neural representation of a concept across the sensory, insular, and motor cortices, and hence operationalises how the perception and action systems provide distributed grounding for words. Each concept is represented as a point (or vector) in an 11-dimensional space of distributed sensorimotor experience, and distance between concepts can therefore be calculated as the distance between the vectors. For example, *alligator* and *monster* are relatively close in sensorimotor terms (i.e., both are experienced primarily by sight, moderately by hearing and head action, weakly by touch and hand action; but are not generally smelled or involve action with the mouth, foot, or torso), whereas *alligator* and *daydream* are quite distant because they share little sensorimotor experience.

Sensorimotor distance is therefore a grounded measure of semantic similarity that operationalises how the distributed neural representations of two concepts across perception

² Some recent distributional models incorporate co-occurrence of either visual or auditory information as well as linguistic (e.g. Bruni, Tran & Baroni, 2014; Lazaridou, Pham & Baroni, 2015; Lopopolo & van Miltenburg, 2015; Günther et al., 2020), but these models fall outside the linguistic domain of the distributional hypothesis, and they are not currently widely used in cognitive psychology to approximate semantic similarity.

and action systems differ from one another³. Its coverage is excellent, as the nearly 40,000 concepts in the Lancaster Sensorimotor Norms is large enough to approximate a full adult conceptual system, covering abstract as well as concrete concepts and all parts of speech, and yielding nearly 800 million comparable concept pairs.

Sensorimotor distance constrains similarity by perception and action experience, and by its nature would also constrain by synonymy (i.e., synonyms like *sofa* and *couch*, or *large* and *big*, would be expected to have extremely similar profiles of sensorimotor experience). Recent work in our lab also suggests that sensorimotor distance appears to capture taxonomic/categorical constraints. For instance, sensorimotor distance between category name and member concept has been successfully used to predict responses in category production (e.g., list as many types of *animal* as you can: Banks et al., 2021) and category verification tasks (e.g., is the pictured *dog* a member of the category *animal*?: van Hoef et al., 2019). Participants were more likely to list a member concept as belonging to a category, and to verify its membership quickly and accurately, when it had short sensorimotor distance from the category concept (e.g., *animal* and *dog*) compared to longer sensorimotor distance (e.g., *animal* and *snake*). Nonetheless, sensorimotor distance would not generally capture all forms of semantic similarity, such as those based on thematic relationships between concepts (e.g., *bee* and *honey*; *grape* and *vineyard*).

³ We note that sensorimotor grounding of word meaning can occur not only via direct, first-hand experience, but also indirectly via vicarious experience or inference from linguistic associations (Barsalou, 1999; Connell & Lynott, 2014; Harnad, 1990; Louwerse, 2011). Because each concept's vector operationalises grounding in sensorimotor systems, regardless of how this grounding is acquired, it means that sensorimotor distance between concepts is also fully grounded in this way. Nonetheless, even though we use vector comparison as the basis for sensorimotor distance, we do not suggest that the mental processes underlying semantic similarity judgements actually resemble vector operations (Jones et al., 2015).

In the current paper, we present the details of the sensorimotor distance measure, and demonstrate that sensorimotor distance has comparable explanatory power to WordNet distance, feature overlap, and CBOW in modelling human similarity judgements while explaining variance in human judgements that is missed by other measures. Furthermore, it does so with the advantages of remaining both grounded and computationally efficient (i.e. easy to calculate via economical representations, once the relevant sensorimotor ratings have already been collected), and applies to both abstract and concrete concepts. All data, analysis code, and full results are available in supplemental materials at <https://osf.io/d42q6/>. We further introduce a web-based tool (available at <https://lancaster.ac.uk/psychology/smdistance>) for easily calculating and visualising sensorimotor distance between lists of concepts, featuring coverage of nearly 800 million concept pairs.

Calculating Sensorimotor Distance

Materials

We took all 39,707 concepts from Lynott et al.'s (2020) Lancaster Sensorimotor Norms, which provide ratings along 11 dimensions of sensorimotor experience as well as a number of other related variables. Lynott et al. normed perceptual and action dimensions separately on a total of 3,500 native speakers of English. For the perceptual norming ($N = 2635$), participants were asked to rate on a scale from 0 (not at all) to 5 (greatly) to what extent they experienced a concept by seeing, hearing, feeling through touch, sensations inside the body, smelling, and tasting (six perceptual modalities, randomly ordered). For the action norming ($N = 1933$), participants were asked to rate on the same scale to what extent they experienced a concept by performing an action with the hand/arm, foot/leg, head excluding mouth, mouth/throat, and torso (5 action effectors, each accompanied by a body avatar image for clarity, randomly ordered). Participants could select a "don't know" button instead of

providing ratings when they were not familiar with the named concept. The final dataset comprised 12.3 million individual ratings and showed excellent inter-rater reliability for all dimensions (Cronbach's alpha = .85–.96). We use here the main form of the norms at the item level, which comprise mean ratings per dimension for 39,707 concepts.

Measures of Sensorimotor Distance

To compute sensorimotor distance between a pair of concepts, we use the vectors of ratings in each of the 11 dimensions of sensorimotor experience. Many possible measures exist for calculating the distance between vectors; here we present *cosine distance* (i.e. 1 minus the cosine of the angle between the vectors⁴), which we found to be the best for modelling human similarity judgements. We also tested four other examples: correlation, Euclidean, Minkowski-3, and Mahalanobis distances⁵, with details included in supplementary materials. Any pair of concepts in the Lancaster Sensorimotor Norms can be compared using cosine distance, yielding sensorimotor distance scores for over 788 million unique concept pairs.

Sensorimotor Distance Characteristics

Sensorimotor distance computations between concept pairs, and other associated functions such as finding nearest neighbours and plotting two-dimensional visualisations, can be performed using an online tool at <https://lancaster.ac.uk/psychology/smdistance>, detailed in Appendix A.

⁴ We opted to use cosine distance rather than cosine similarity for consistency with other distance measures in the web tool and to avoid terminological confusion with semantic similarity.

⁵ While Lynott et al. (2020) discussed some variables with similar names in the original Lancaster Sensorimotor Norms paper, they all compressed the multidimensional sensorimotor strength of *a single concept* by calculating the distance from a concept vector to the origin. By contrast, the measures we present here compare *two separate concepts* by calculating the distance between their individual vectors.

Distance distributions. Cosine distances between non-negative vectors range in theory from 0 to 1, and sensorimotor distance measures span almost the entire range of possible values: the minimum attained distance is .0002 (the closest pair is *cyan–pixilation*, with other very close pairs including *hyphen–colorfast*, distance 0.0020, and *everything–multisensory*, distance 0.0038; excluding the distances of zero between each concept and itself) and the maximum is .950 (the furthest pair is *shinbone–smelled*, with other very distant pairs including *flavorless–handgrip*, distance 0.942, and *adobe–digestion*, distance 0.921). The full distribution of distances is shown in Figure 1. Mean sensorimotor distance was .195 ($SD = .123$).

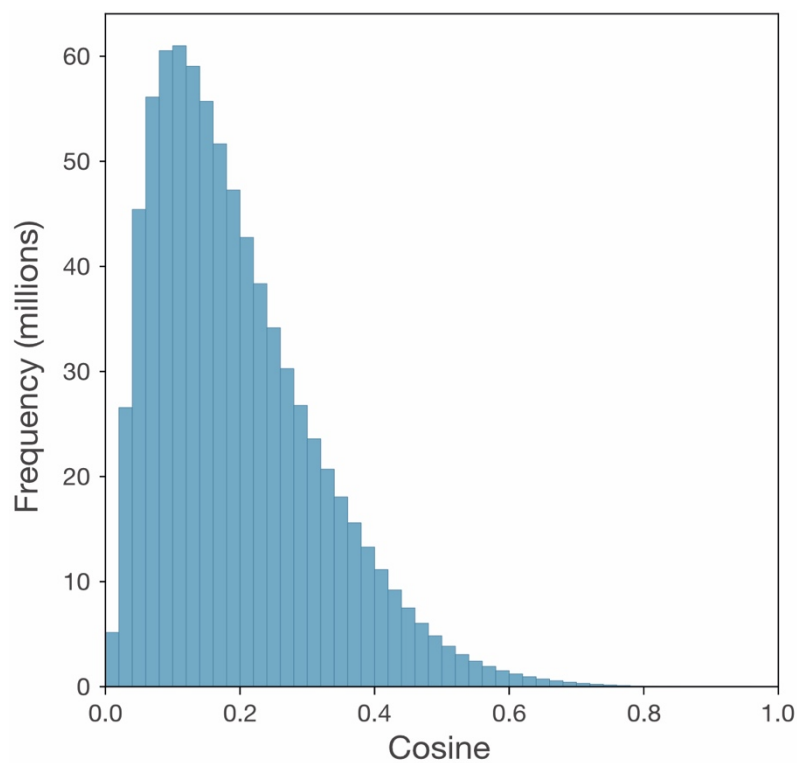


Figure 1. Distributions of cosine distances between all (approximately 800 million) pairs of concepts in the Lancaster Sensorimotor Norms.

Visualizing distance between concepts. The relative distances between select concepts can be visualized using multidimensional scaling (MDS) techniques, which arrange points in two-dimensional space while minimizing the distortion of the pairwise distances. Figure 2 shows two examples of such MDS plots for a selection of category exemplars taken from the norms, demonstrating clustering between semantic categories of nouns and action categories of verbs (see also Connell et al., 2019).

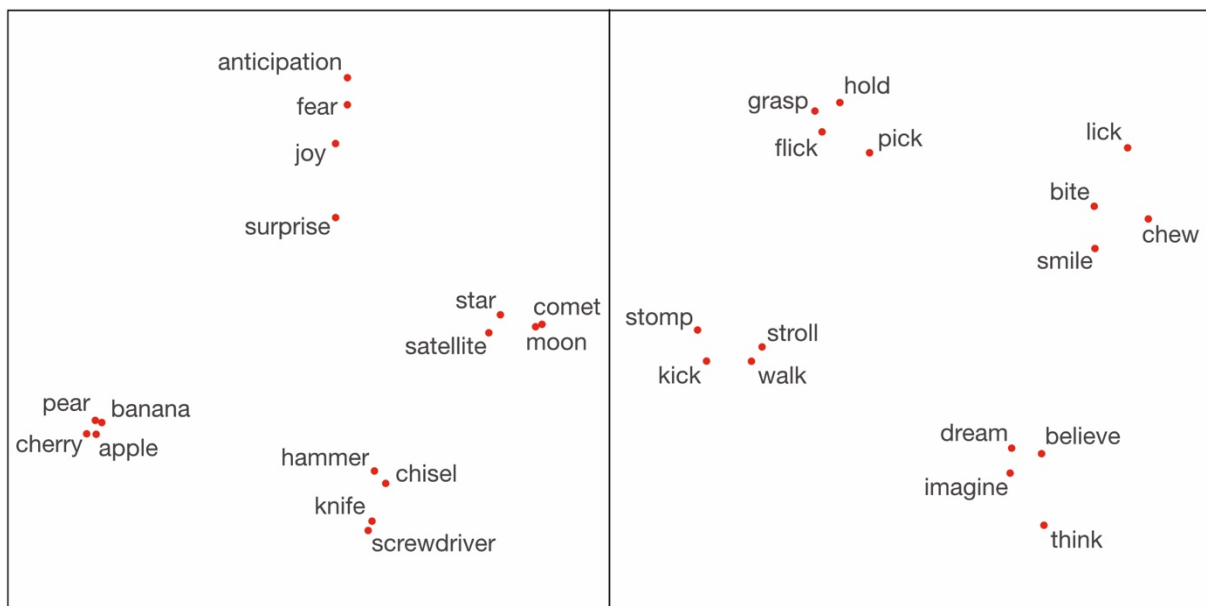


Figure 2. Visualizing sensorimotor distance between sample concepts. Cosine distances between each pair of concepts were transformed using nonmetric multidimensional scaling (Sammon, 1969). Left panel: select nouns for tools, emotions, fruit and celestial objects. Right panel: select verbs for leg, hand, mouth and cognitive actions.

Nearest neighbours. From a reference word, lists of nearest sensorimotor neighbours (i.e. the other concepts which have the smallest distance to the reference word) can be generated. Some examples of nearest neighbours are shown in Figure 3, suggesting that sensorimotor distance can encode detailed information about concepts (e.g. speed of movement).

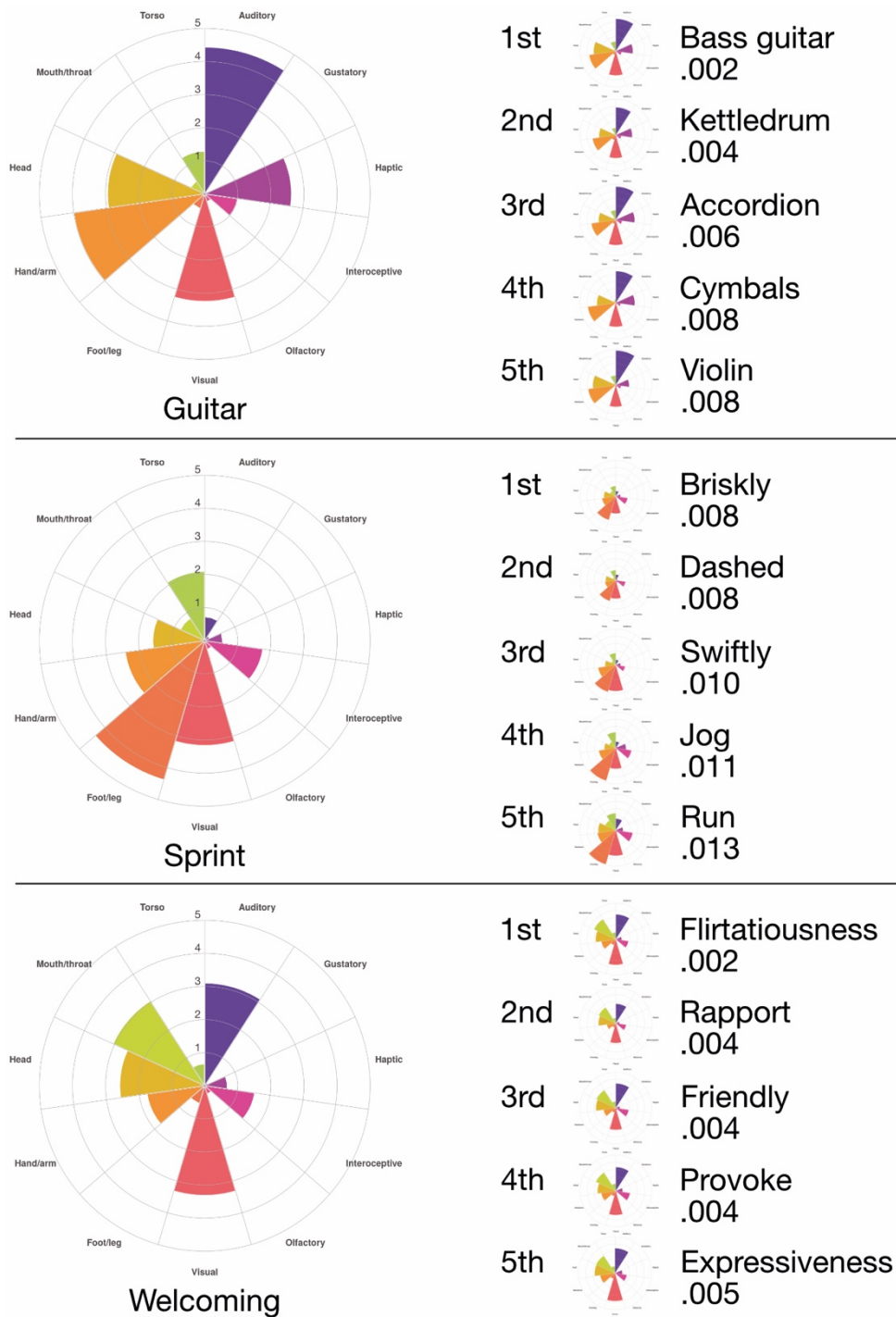


Figure 3. Examples of top-5 nearest neighbors in sensorimotor distance. Each concept is accompanied by a polar plot, which shows the strength of rating in each dimension: (clockwise from the top) auditory, gustatory, haptic, interoceptive, olfactory, visual, foot/leg, hand/arm, head, mouth/throat, torso.

Validating Sensorimotor Distance

For sensorimotor distance to be a useful research tool, it is important to show both how it compares to other measures of semantic similarity, and that it is a good predictor of human judgements of similarity which are missed by other measures. All materials, data and associated statistics are available in supplementary materials at <https://osf.io/d42q6/>.

Analysis 1: Comparison to Other Measures of Semantic Similarity

In this first analysis of convergent validity, we compare sensorimotor distance as a grounded measure of semantic similarity with alternative similarity measures that originate in different theoretical perspectives on the conceptual system: hierarchical structure (i.e., WordNet distance), feature-based representations (i.e., feature overlap), and linguistic distributional information (i.e., CBOW). Overall, sensorimotor distance correlates as well with alternative measures of semantic similarity as such measures do with each other.

Method & Materials. We compiled 4,325 word pairs featured in existing datasets of human similarity ratings: WordSim (Finkelstein et al., 2002), Simlex (Hill et al., 2016) and MEN (Bruni et al., 2014). Coverage varied by measure, as outlined below.

As well as our own sensorimotor distance measure, we selected three popular measures of semantic similarity which have been widely used across the cognitive sciences, each relating to one of the theoretical frameworks earlier discussed:

Sensorimotor distance. A total of 3,730 word pairs were covered by our database, for which we calculated sensorimotor distance (cosine distance $M = .126$ $SD = .104$).

WordNet distance. Maki et al. (2004) compared several related measures based on distance in the WordNet taxonomy, from which the authors determined that Jiang–Conrath distance (Jiang & Conrath, 1997) to be the superior choice for modelling semantic similarity. Jiang–Conrath distance similarity is based on the information content of two concepts relative to that of their most specific mutual ancestor in the hierarchy (i.e. the “least common

subsumer”). Although Maki et al. make available a database of precomputed distances for around 50,000 word pairs, it covered only approximately 10–15% of most of the similarity datasets of we set out to model here. We therefore opted to recompute Jiang–Conrath distances on WordNet using the implementation in NLTK version 3.2 (Bird et al., 2009), which covered 3,776 word pairs (WordNet distance $M = 11.39$, $SD = 6.06$).

Feature overlap. Buchanan et al. (2019) collected feature-production norms for a list of 4,436 concepts. Pairs of concepts can be compared via their respective lists of norms⁶. Instead of counting the number of norms in common between a concept pair, Buchanan et al. recommend computing the cosine of the angle between the sparse property-frequency vectors (yielding approximately 10 million comparable pairs). Buchanan et al. provide a database of precomputed cosine-overlap values for just over 208,000 pairs, which covered 2,414 pairs from our item set (Feature overlap $M = .095$, $SD = .189$).

CBOW. The computation of CBOW scores involves training a neural network model on a huge corpus of text to predict a target word from its linguistic contexts (Mikolov et al., 2013). We used the CBOW vectors from Mandera et al. (2017, provided by Mandera, 2016) to calculate cosine distances for our materials: 4,325 word pairs were covered (CBOW cosine distance $M = .693$, $SD = .162$).

Analysis. We computed Bayesian correlations between all four semantic similarity measures using JASP (JASP Team, 2020) with a stretched prior beta width = 1 (i.e., uniform prior where all correlations values are equally likely). Because some similarity measures were distances (i.e., more similar = lower score) while others were similarity/overlap scores (i.e., more similar = higher score), the direction of the alternative hypothesis varied. Matching

⁶ Buchanan et al. (2019) provide two measures of feature overlap using cosine distance: one using the original features provided by their participants ("raw" overlap), and another where features were combined using a root lemmatiser ("root" overlap). In what follows we use the "root" overlap measure.

constructs were expected to be positively correlated (i.e., between sensorimotor distance, WordNet and CBOW), whereas mismatching constructs (i.e., all other comparisons) were expected to be negatively correlated. Bayes Factors (BF) are reported as natural logarithms due to their magnitude.

Results. Sensorimotor distance correlated at best moderately with other measures of semantic similarity (see Figure 4), with very strong evidence that the correlations ran in the expected direction: all log BFs > 80 . Intercorrelations between WordNet distance, feature overlap, and CBOW scores were of similar magnitude, indicating that sensorimotor distance correlated with other measures of semantic similarity about as well as they correlate with each other. Full statistics for all comparisons can be found in supplementary materials.

Sensorimotor distance therefore incorporates unique information that is not captured by other measures of semantic similarity, although it is not yet clear whether this unique information reflects semantic similarity itself as opposed to mere noise. We address this question in the following section by examining its external validity in predicting human similarity judgements.

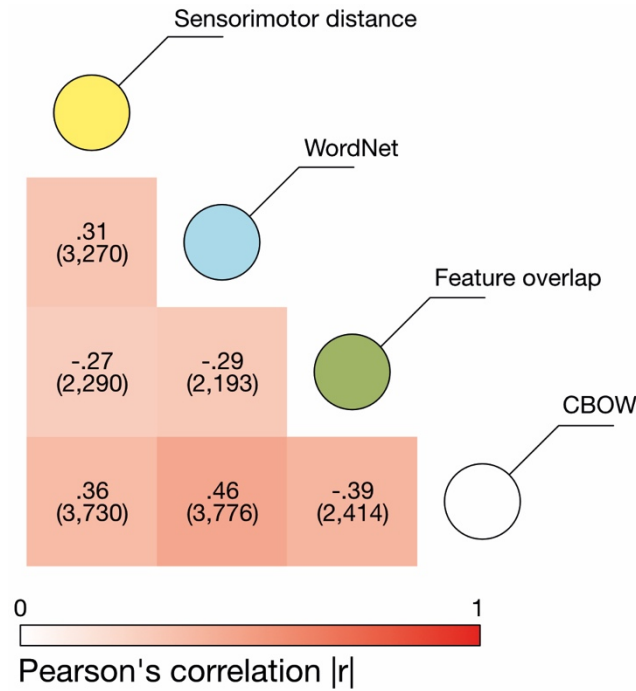


Figure 4. Correlations between sensorimotor distance and three other measures of semantic similarity. The color scale indicates the absolute value of the correlation (i.e., stronger color = stronger relationship) while the correlation sign varies according to whether the variable is a measure of distance or similarity/overlap. The number of pairs per comparison is in parentheses (2,081 pairs were common to all measures).

Analysis 2: Predicting Human Similarity Judgements

In this section, we demonstrate external validity by examining how effectively sensorimotor distance can predict human judgements of semantic similarity and compare its performance to other measures. Using three different datasets of human similarity judgements, we demonstrate that sensorimotor distance can explain unique variance above and beyond each alternative measure of semantic similarity (i.e., WordNet, feature overlap, CBOW). In addition, given that each measure constrains semantic similarity in a different way that is potentially useful to modelling human data, we examine what combination of

semantic similarity measures best explains human similarity judgements. Across the three datasets of human similarity data, we find that sensorimotor distance is consistently included in the best-fitting model and demonstrates the most consistent level of performance.

Method & Materials. To compare the relative explanatory power of each model of semantic similarity, we examined participant similarity judgements from three existing datasets: Simlex-999 (Hill, n.d.; Hill et al., 2016: 999 word pairs), WordSim-353 (Gabrilovich, 2002; Finkelstein et al., 2002: 353 word pairs), and MEN (Bruni, 2012; Bruni et al., 2014: 3,000 word pairs). In the Simlex and WordSim datasets, participants directly rated the similarity of pairs of words and the dependent variable is the mean similarity rating per word pair. In the MEN dataset, however, participants selected the most closely related out of two possible word pairs in a forced-choice paradigm; these choices were then converted into a single similarity score for each pair. From each dataset, we selected only those items that were covered by all four of the semantic similarity measures, resulting in 669 word pairs from Simlex, 181 from WordSim, and 1,251 word pairs from MEN.

Analysis. Each dataset was analysed separately but identically in three stages. We first computed zero-order correlations between the human similarity scores and each of the four semantic similarity measures (i.e. sensorimotor distance, WordNet distance, feature overlap, CBOW); Bayesian correlation were carried out in JASP as per previous section.

Next, to examine the independent contribution of sensorimotor distance, we carried out hierarchical Bayesian linear regressions (JASP Team 2020: using JSZ default priors, r scale = .354, beta binomial distribution $a = 1$ and $b = 1$) on human similarity judgements. Step 1 entered one of the other semantic similarity measures (i.e., WordNet distance, feature overlap, or CBOW scores), and Step 2 entered sensorimotor distance. Model comparisons using Bayes Factors (BF) between steps therefore tested whether sensorimotor distance explained unique variance in human similarity judgement above and beyond other similarity

measures. In this analysis, log BF for Step 2 over Step 1 is equivalent to the inclusion Bayes Factor (BF-inclusion)

Finally, to find the best possible model of human similarity judgements for each of the three datasets, we conducted Bayesian linear regressions (settings as above) by examining all possible combinations of all four semantic similarity measures as predictors, and selecting the model that offered the best fit to that dataset. We also report inclusion Bayes Factors (BF-inclusion) for each predictor, which reflects the change from prior to posterior odds for all models including a particular predictor compared to models excluding it (Hinne et al., 2020), and allows us to compare the relative strength of evidence for each similarity measure in predicting each dataset of human similarity judgements.

Results. Figure 5 shows zero-order correlations between each semantic similarity measure and human similarity judgements from each dataset. Sensorimotor distance was moderately correlated with human similarity scores (i.e., shorter distance = more similar), with the magnitude of the correlations within the bounds achieved by alternative similarity measures. All correlations had very strong evidence in the expected direction (log BF_s > 13.4).

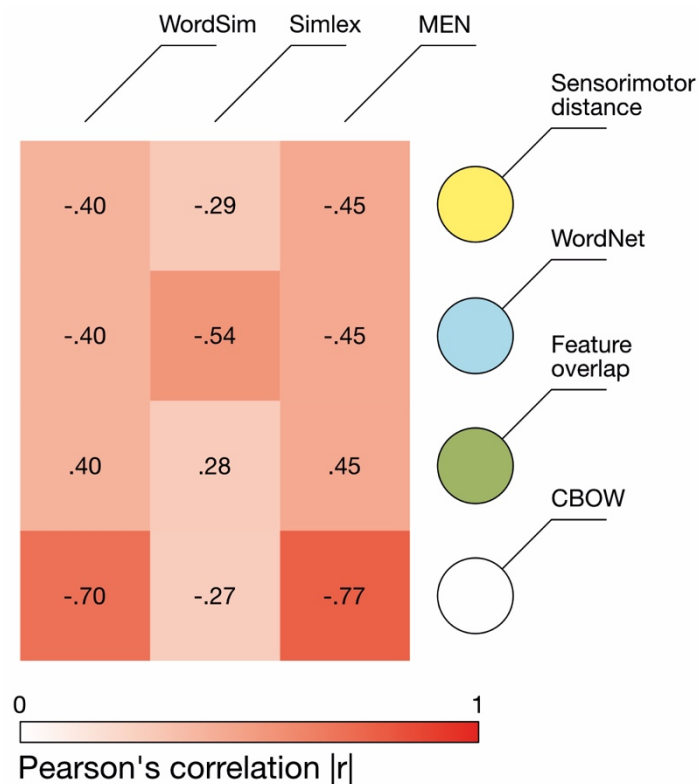


Figure 5. Zero-order correlations between human similarity judgements and each measure of semantic similarity, calculated separately per dataset. The color scale indicates the absolute value of the correlation (i.e., stronger color = stronger relationship) while the correlation sign varies according to whether the variable is a measure of distance or similarity/overlap.

In the hierarchical regression analyses, there was strong evidence for the inclusion of sensorimotor distance at Step 2 in all models: see Figure 6 for change in R^2 and Table 1 for coefficients. For all three datasets, sensorimotor distance explained variance in human similarity judgements above and beyond that explained by alternative measures of semantic similarity (i.e., WordNet distance, feature overlap, CBOW). In all analyses, variance inflation factors were approximately 1, indicating that multicollinearity was not an issue.

Overall, these results indicate that the unique information captured by sensorimotor distance is *not* mere noise. Rather, they suggest that sensorimotor distance constrains similarity in a way that is not captured by other measures of semantic similarity that relate to hierarchical structure (i.e., WordNet distance), feature-based representations (i.e., feature overlap), or linguistic distributional information (i.e., CBOW).

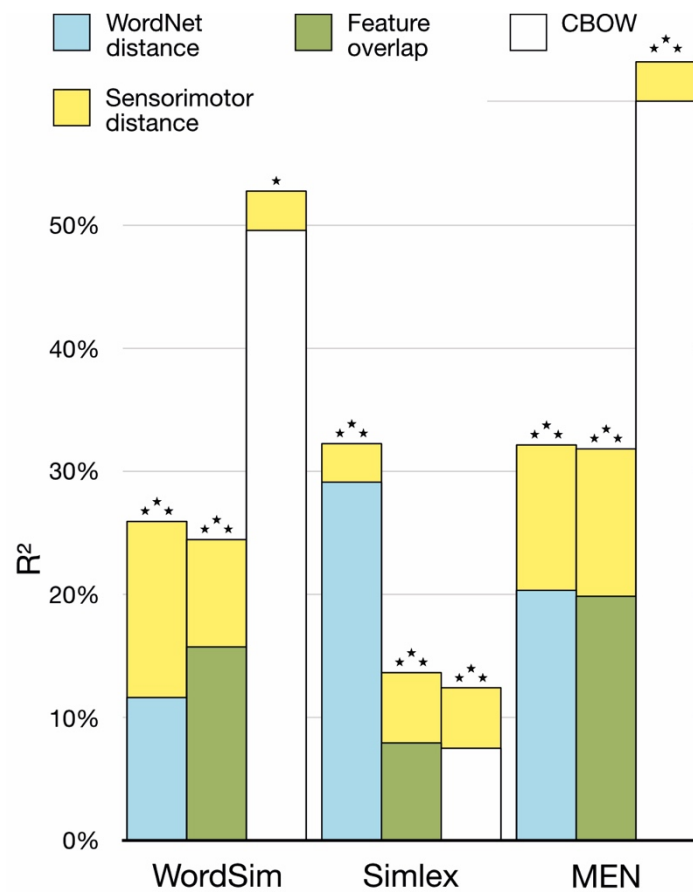


Figure 6. Unique effects of sensorimotor distance (top of stacked bar; yellow) in explaining variance in human similarity judgements when added to a regression model already containing an alternative measure of semantic similarity (bottom of stacked bar; color varies). Regressions were performed separately for each dataset (WordSim, Simlex, MEN) and alternative measure of semantic similarity (WordNet distance, feature overlap, CBOW). Asterisks indicate evidence for including sensorimotor distance at Step 2 compared to alternative predictor at Step 1 (* log $BF_{10} > \log 10$; ** log $BF_{10} > \log 100$; *** log $BF_{10} > \log 1000$).

Table 1.

Regression coefficient statistics for Step-2 models of human similarity judgements across three datasets, showing coefficient estimates and their 95% credible intervals for sensorimotor distance and each alternative semantic similarity measure, as well as natural log of inclusion Bayes Factors for the sensorimotor distance predictor.

Dataset	Step-1 semantic predictor	Step-2 sensorimotor distance coefficient	95% CI		Log BF-inclusion
			Lower	Upper	
WordSim	WordNet	-7.30	-10.60	-4.69	9.09
	Feature overlap	-6.90	-10.31	-4.21	7.45
	CBOW	-4.35	-7.19	-1.63	3.37
Simlex	WordNet	-5.13	-6.96	-3.35	12.24
	Feature overlap	-6.84	-8.89	-4.84	18.64
	CBOW	-6.50	-8.61	-4.44	15.62
MEN	WordNet	-39.39	-44.65	-34.23	96.72
	Feature overlap	-39.71	-44.97	-34.53	97.99
	CBOW	-21.59	-25.54	-17.59	49.49

Finally, in the best-model regressions, the optimal predictors of human similarity judgement varied by dataset, as did the relative evidence for each predictor (see Figure 7 for summary and Table 2 for coefficient statistics). For WordSim, only CBOW and sensorimotor distance were included in the best model, which explained over half the variance with a very strong level of evidence ($R^2 = .527$, $\log BF_{10} = 60.92$; full statistics for all candidate models are available in supplemental materials). Inclusion Bayes Factors indicated that CBOW was the best predictor of human similarity judgements in WordSim, followed by sensorimotor distance. Notably, there was evidence *against* including feature overlap as a predictor of

human similarity judgements in WordSim (i.e., a model including feature overlap was $\log BF_{10} = -2.14$ times worse than the best model of just CBOW and sensorimotor distance), and no positive evidence for including WordNet distance (i.e., a model including WordNet distance was $\log BF_{10} = -0.76$ times worse than the best model). For Simlex, the best model comprised (in rank order of BF-inclusion) Wordnet distance, sensorimotor distance, and feature overlap, which explained a third of the variance with a very strong level of evidence ($R^2 = .338$, $\log BF_{10} = 128.12$). In this case, there was evidence against including CBOW as a (i.e., a model containing all four measures was $\log BF_{10} = -2.383$ times worse than the best model that excluded CBOW), despite it being the best predictor of WordSim similarity. For the MEN dataset, all measures of semantic similarity were included in the best model, which this time explained a very high 65% of variance with a very strong level of evidence ($R^2 = .651$, $\log BF_{10} = 642.08$). The best predictor by BF-inclusion was CBOW, followed by sensorimotor distance, then feature overlap, and lastly WordNet distance (i.e., the weakest predictor of MEN similarity despite being the best predictor of Simlex similarity).

Overall, these best-model regressions showed that no single measure of semantic similarity was consistently preferred as the top predictor of human similarity judgements. Sensorimotor distance was present in every best model, and no other predictor was consistently ranked better across all datasets. On the other hand, sensorimotor distance was never the overall best predictor, and was only consistently preferred to feature overlap over all datasets. We note that the pattern of results changed little when we examined an alternative linguistic distributional model (LSA; see Appendix B), which suggests that our findings generalise beyond the particular implementation of CBOW (e.g., corpus size can affect performance: see Bullinaria & Levy, 2012; Wingfield & Connell, 2022). This pattern of results is consistent with the idea that different measures of semantic similarity constrain

similarity in different ways, all of which are relevant to what humans consider when judging the similarity of concepts.

Table 2.

Regression coefficient statistics for the most complex model of human similarity judgements across three datasets, showing coefficient estimates and with 95% credible intervals for sensorimotor distance and each alternative semantic similarity measure, as well as natural log of inclusion Bayes Factors for each predictor.

Dataset	Parameter	Coefficient	95% Credible Interval		Log BF-inclusion
			Lower	Upper	
WordSim	Intercept	5.88	5.64	6.07	
	WordNet	-0.04	-0.07	1.49E-4	-0.12 ^a
	Feature overlap	0.20	-0.43	1.41	-1.17 ^b
	CBOW	-7.44	-9.35	-6.20	35.62
	Sensorimotor distance	-4.00	-7.19	-1.36	3.24
Simlex	Intercept	4.34	4.18	4.49	
	WordNet	-0.22	-0.25	-0.19	80.19
	Feature overlap	1.47	0.69	2.19	5.54
	CBOW	0.07	-0.57	0.92	-1.00 ^c
	Sensorimotor distance	-4.54	-6.39	-2.74	9.71
MEN	Intercept	25.45	25.04	25.85	
	WordNet	-0.14	-0.25	-0.05	3.45
	Feature overlap	9.46	6.88	12.00	23.43
	CBOW	-45.93	-49.43	-43.40	341.18
	Sensorimotor distance	-18.87	-22.83	-14.97	40.33

^a Indicates equivocal evidence *against* inclusion of WordNet scores in model of WordSim dataset. ^b Indicates evidence *against* inclusion of feature overlap scores in model of WordSim dataset. ^c Indicates equivocal evidence *against* inclusion of CBOW scores in model of Simlex dataset

Dataset	Predictors				
	1 (Best)	2	3	4 (Weakest)	Not a predictor
WordSim	○ CBOW	● Sensorimotor distance			● Feature overlap ○ WordNet
Simlex	● WordNet	● Sensorimotor distance	● Feature overlap		○ CBOW
MEN	○ CBOW	● Sensorimotor distance	● Feature overlap	● WordNet	

Figure 7. Rank order from best to worst of each semantic similarity measure in predicting human similarity judgements across three datasets, based on inclusion Bayes Factors in best-model regressions.

Sensorimotor Distance for Abstract and Concrete Concepts

As a measure of semantic similarity that is based on perception and action experience, some might wonder whether sensorimotor distance could apply to abstract concepts, which in some accounts are defined by their lack of perceptual information (e.g., Paivio, 1986). Previous research has shown that virtually all concepts, regardless of their concreteness, are experienced to some extent through various sensorimotor dimensions. Connell & Lynott (2012) showed that many abstract concepts tend to be strongly perceptual (i.e., their experience involves perception, particularly vision), Connell et al. (2018) found that interoceptive strength (i.e., sensations inside the body) was *more* important to abstract concepts than to concrete, and Lynott et al.’s (2020) norms demonstrate multidimensional sensorimotor profiles for many abstract concepts such as *justice* and *everything*. In principle, therefore, sensorimotor distance should apply as a semantic similarity between abstract concepts as well as between concrete concepts (see also Figure 3).

To examine this principle in action, we compared the ability of sensorimotor distance to predict human similarity judgements in three different categories of concept pairs: both

concepts abstract (e.g., *inexpensive* and *cheap*), mixed concrete–abstract (e.g., *battle* and *conquest*), and both concepts concrete (e.g., *drizzle* and *rain*).

Method & Materials

Of the three datasets of human similarity judgements examined in Validation Analysis 2, only one contained sufficient numbers of abstract concepts to enable meaningful comparisons: Simlex-999 (Hill et al., 2016)⁷. Using Brysbaert et al.'s (2014) concreteness ratings, we categorised concepts as abstract if their rating was < 3 (i.e., the concreteness scale midpoint) and as concrete if their rating was ≥ 3 . Sensorimotor distance was available for 993 of 999 Simlex concept pairs, which we then split as follows: 264 abstract–abstract pairs, 172 mixed pairs (i.e., one abstract, one concrete), and 557 concrete–concrete pairs.

Analysis

We computed Bayesian correlations between sensorimotor distance and Simlex similarity judgements (JASP Team, 2020) with a stretched prior beta width = 1 (i.e., uniform prior where all correlations values are equally likely), and the alternative hypothesis that the variables would be correlated negatively (i.e., more similar = shorter distance). Correlations were computed separately for each category of concept pair.

Results

Sensorimotor distance correlated with human similarity judgement comparably well for all categories of concept pair (see Figure 8). The highest correlation was actually for mixed word pairs, but – importantly – the correlations for abstract–abstract pairs and concrete–concrete pairs were close in magnitude, and comparable given their 95% credible intervals. These results suggest that sensorimotor distance is a useful measure of semantic similarity for all concept pairs; abstract and concrete alike.

⁷ There were 264 abstract–abstract pairs for which sensorimotor information was available (26% of dataset) in Simlex-999, but only 41 in WordSim-353 (11% of dataset) and 17 in MEN (<1% of dataset).

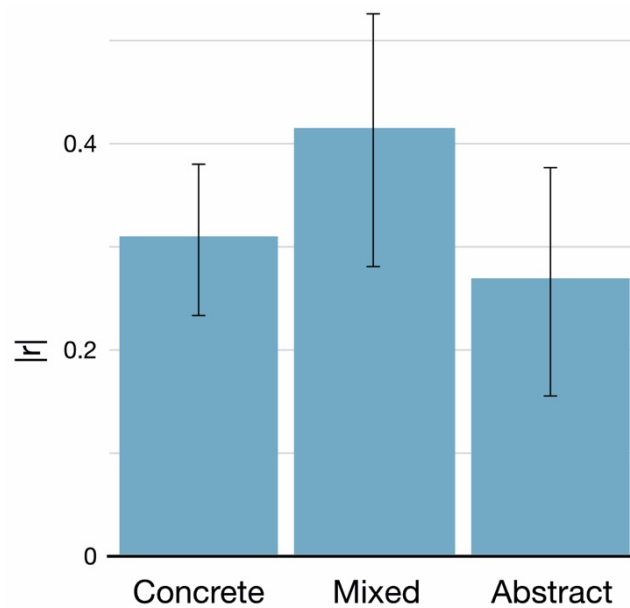


Figure 8. Absolute value of correlations between sensorimotor distance and human similarity judgement for the Simlex dataset, for word pairs where both are concrete, both are abstract, and mixed pairs. Error bars show the 95% credible intervals for the correlation value.

General Discussion

We have presented sensorimotor distance, a novel grounded measure of semantic similarity for nearly 800 million concept pairs that is based on Lynott et al.'s (2020) 40,000-concept sensorimotor strength norms. Unlike existing measures of semantic similarity (e.g. CBOW, WordNet, feature overlap), sensorimotor distance directly operationalises sensorimotor experience in multiple perceptual modalities and action effectors and is therefore grounded in how it constrains similarity. The semantic information represented by sensorimotor distance is transparent, relevant to all concepts/words regardless of their concreteness or grammatical class, and is available at a scale that covers a full-size adult conceptual system for a native speaker of English.

In validating sensorimotor distance, we demonstrated that it captures information about semantic similarity that is not captured by alternative measures, and that human judgements of similarity are best fit by combining multiple similarity measures in a single model. Indeed, the optimal combination of similarity measures varied markedly from one dataset to the next, which highlights the importance of validating semantic similarity measures against multiple human benchmarks, yet sensorimotor distance was the most consistent predictor across datasets. These findings support the idea that, when people judge if things are semantically similar, they employ multiple constraints on what similarity might mean. Multiple measures of similarity, each applying different constraints, are therefore required to fully capture the similarity between two given concepts (see Goodman, 1972).

Like many semantic predictors used in cognitive psychology (including some other predictors used in this study: feature overlap and taxonomic distance), sensorimotor distance is ultimately derived from participant responses in a task which involves access to words' semantic representations. Therefore, insofar as such a predictor is used to model cognitive processes or representations which themselves involve accessing word semantics—as is common in the cognitive sciences—it cannot account for the dereferencing of mental concepts from their labels *per se* (Westbury, 2016; Wittgenstein, 1958). In theory, one might hope to derive the multidimensional vector from direct recordings of activation in participants' sensorimotor cortices (e.g. Hauk et al., 2004) while they experience (and recall, name, etc.) various concepts across various contexts, and to use these recordings to quantify the degree to which different perceptual modalities and action effectors were involved in direct experience of each particular concept. Such measurements – and any resulting distance calculations between concepts – would qualify as an out-of-domain explanation of (part of) word semantics that would satisfy Westbury's (2016) concerns about dormitivity. Of course, in reality, it would be completely impractical to conduct this hypothetical norming study at

the scale of tens of thousands of concepts that comprise the human conceptual system (e.g. Hauk et al. required high-resolution functional and structural MRI scans to localise 14 participants' responses to 150 test words). Instead, the measures that underlie sensorimotor distance (i.e., the Lancaster Sensorimotor Norms), as explained by Lynott et al. (2020), aim to approximate it via introspective judgements of sensorimotor experience. We believe that by restricting the domain of judgement so tightly, the Lancaster Sensorimotor Norms provide a reasonable proxy for direct sensorimotor experience (see also Reilly et al., 2020) in a tractable way, as well as allowing the pool of items to easily extend to traditionally abstract and/or physically diffuse concepts (e.g. *democracy*, which is perhaps easier to characterise through introspection than to experience in a lab setting) that nonetheless appear to have a robust, situated, sensorimotor grounding. Sensorimotor distance, based on this reasonable proxy of sensorimotor experience, therefore provides a tractable operationalisation of how the distributed representations of two concepts across perception and action systems differ from one another.

Of course, the particular 11 dimensions that we use here to calculate sensorimotor distance are not the only possible way to specify dimensions of perception and action experience. Although each dimension is well motivated (see Lynott et al., 2020, for details), they exhibit a complex intercorrelational structure that corresponds to how the human body's senses and effectors interact with the external world. This structure reflects, for example, that things which can be touched can usually be seen, or that things which can be tasted can usually also be smelled but are not usually subject to action with the foot/leg. As a result, one might be concerned that some dimensions are redundant, and that cosine distance therefore produces a skewed picture of what sensorimotor distance should reflect. However, cosine distance (which is sensitive to this correlated structure) overall *outperforms* Mahalanobis distance (which removes this correlated structure: Mahalanobis, 1936; see supplementary

materials for full results), which suggests that the present 11-dimensional space is a reasonably accurate reflection of how sensorimotor information informs human judgements of semantic similarity. Nonetheless, Mahalanobis distance is available in the web tool for researchers who explicitly wish to use it⁸. Alternatively, one may wonder if more fine-grained distinctions of sensorimotor experience would be useful, so long as they still meet the same criteria as the original dimensions (i.e., perception or action experience that is processed in a distinct cortical region). For example, visual perception could be subdivided into colour versus visuospatial movement, haptic perception could be subdivided into sensation on the hand versus elsewhere on the body, hand/arm action could be subdivided into action of the hand versus the arm/shoulder area, and so on. Whether such fine-grained distinctions would help or hinder the accuracy of sensorimotor distance in predicting semantic similarity remains an open question for future research.

We hope that sensorimotor distance, available in an online application at <https://lancaster.ac.uk/psychology/smdistance> (see Appendix A), will provide a useful tool for researchers in cognitive psychology, psycholinguistics, cognitive neuroscience, or any field relevant to semantic similarity and the grounded nature of concepts in semantic memory.

Conclusion

We hope that sensorimotor distance, available in an online application at <https://lancaster.ac.uk/psychology/smdistance> (see Appendix A), will provide a useful tool for researchers in cognitive psychology, psycholinguistics, cognitive neuroscience, or any field relevant to semantic similarity and the grounded nature of concepts in semantic memory.

⁸ We thank Fritz Günther for this suggestion

Appendix A: Web tool for sensorimotor distance calculation

For convenience, sensorimotor distances and related operations can be computed using a web-based tool developed by the authors and available at <https://www.lancaster.ac.uk/psychology/smdistance/>. The available functions include calculating sensorimotor distance between concepts (pairwise, one-to-many, or many-to-many matrix), producing a list of nearest neighbors in sensorimotor space for a given concept, and visualizing 2-dimensional representations of sensorimotor distance between concepts. Figure A1 illustrates some examples of the interface.

Calculate distances between concepts:
One-to-one

Calculate distances between pairs of concepts' vector representations.

Concept pairs

morocco : cloning
instantaneousness : puritanical
tolerant : distiller
unrevised : forklift
lactic : loathsome
upkeep : symptomatically
operate : interdependent
changer : quantify
squeaky : crustacean
grind : subspecialty

Enter pairs of concepts separated by colons, commas, tabs, or semicolons. Enter each pair on a separate line.

Clear 10 valid pairs entered.

Word 1	Word 2	Cosine distance
morocco	cloning	0.117670
instantaneousness	puritanical	0.064573
tolerant	distiller	0.245347
unrevised	forklift	0.122908
lactic	loathsome	0.327242
upkeep	symptomatically	0.167177
operate	interdependent	0.198870
changer	quantify	0.061765
squeaky	crustacean	0.593757
grind	subspecialty	0.134741

Distance computation (pairwise)

Calculate distances between concepts:
Many-to-many (disatnce matrix)

Calculate distances between the vector representation of different concepts.

Maximum of 200 items per list. If you require more items, please consider **downloading the full dataset** and computing pairwise distances from there.

Concepts

inert
wavering
barf
devalue
tap water
accuracy
cheeseburger
unsigned
platter
nether

	inert	wavering	barf	devalue
inert	0.000000	0.229781	0.247530	0.229788
wavering	0.229781	0.000000	0.253324	0.141091
barf	0.247530	0.253324	0.000000	0.181513
devalue	0.229788	0.141091	0.181513	0.000000
tap water	0.239530	0.233148	0.152445	0.179720
accuracy	0.181992	0.117675	0.270434	0.092860
cheeseburger	0.228997	0.405437	0.167248	0.377069
unsigned	0.168562	0.248696	0.468768	0.229844
platter	0.143950	0.228160	0.307909	0.195988
nether	0.224384	0.085487	0.318158	0.162357

Distance computation (matrix)

Find neighbours

Find nearest neighbours of a concept via its vector representation.

Concept

shower gel

Enter a concept here.

Clear

Number

10

Limit to this many nearest neighbours.

Within distance

Any distance

Limit to neighbours within a fixed distance.

Nearest neighbours of "shower gel"

Download [.csv]

Order	Concept	Cosine distance
1	soap	0.007669
2	suntan lotion	0.008140
3	lotion	0.008332
4	moisturizer	0.016480
5	body lotion	0.016646
6	sunscreen	0.019705
7	gardening	0.024167
8	fertilize	0.025018
9	toiletry	0.028049
10	housecleaning	0.034013

Nearest neighbours

Visualise concepts

Plot an MDS (multidimensional scaling) arrangement of concepts using their vector representations. Please enter a minimum of 3 and a maximum of 20 items.

For Euclidean and Minkowski-3 distances, classical metric MDS is used. For cosine and correlation distances, Sammon nonmetric MDS is used.

Concepts

grasp
hold
flick
pick
stomp
kick
stroll
walk
lick
bite

Two-dimensional visualisation

Figure A1. Four sample modes of the Sensorimotor Distance online interface, showing (clockwise from top left): computing distance between pairs of words; computing distance matrices between lists of words; visualizing sensorimotor distance between concepts in 2 dimensions; finding nearest neighbors.

Appendix B: A comparison with latent semantic analysis

Latent semantic analysis (LSA: Landauer & Dumais, 1997) continues to have extremely widespread use in the cognitive sciences as a measure of semantic similarity (e.g. Dautriche et al., 2017; Gagné et al., 2020; Ren & Coutanche, 2021). We examined LSA as an alternative linguistic distributional model to CBOW. Overall, LSA was a somewhat worse predictor of human similarity judgements than CBOW, but it exhibited the same qualitative patterns. Critically, it exhibited a very similar pattern relative to the other predictors, and in particular relative to sensorimotor distance. We include the relevant figures of results below; full details are available in supplemental materials⁹ (<https://osf.io/d42q6/>).

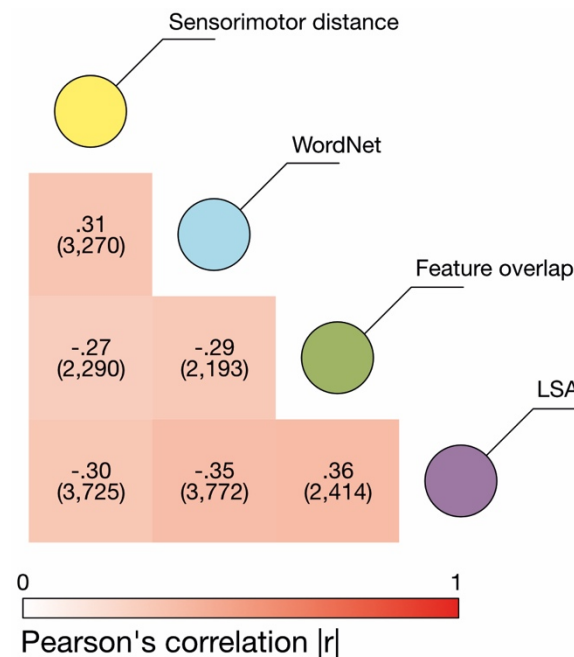


Figure B1. Correlations between sensorimotor distance and three other measures of semantic similarity. The color scale indicates the absolute value of the correlation (i.e., stronger color = stronger relationship) while the correlation sign varies according to whether the variable is a measure of distance or similarity/overlap. The

⁹ In a separate analysis we found that both perceptual and action domains contribute to sensorimotor distance effects. Results of this analysis can also be found in the supplementary materials. We thank an anonymous reviewer for this suggestion.

number of pairs per comparison is in parentheses (2,081 pairs were common to all measures). (See Figure 4 for comparison with CBOW.)

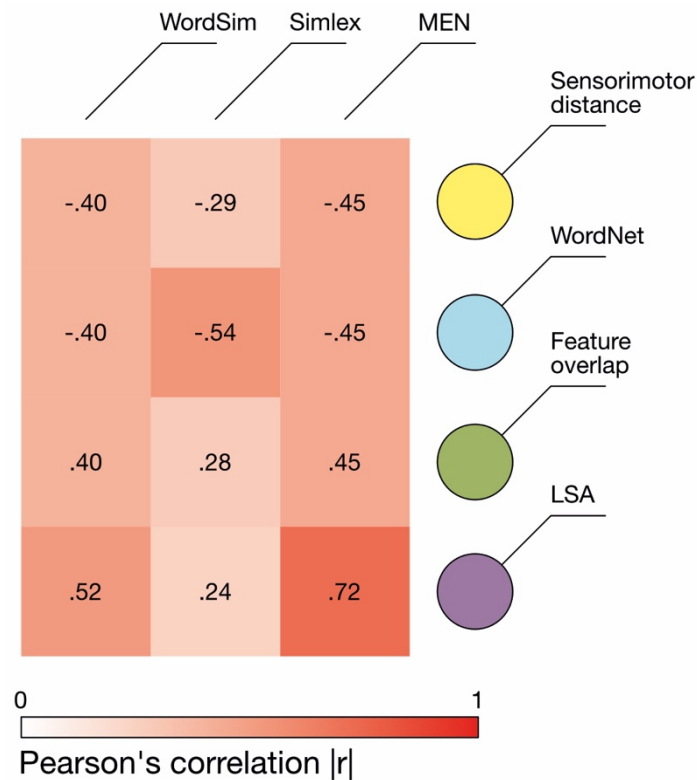


Figure B2. Zero-order correlations between human similarity judgements and each measure of semantic similarity, calculated separately per dataset. The color scale indicates the absolute value of the correlation (i.e., stronger color = stronger relationship) while the correlation sign varies according to whether the variable is a measure of distance or similarity/overlap. (See Figure 5 for a comparison with CBOW.)

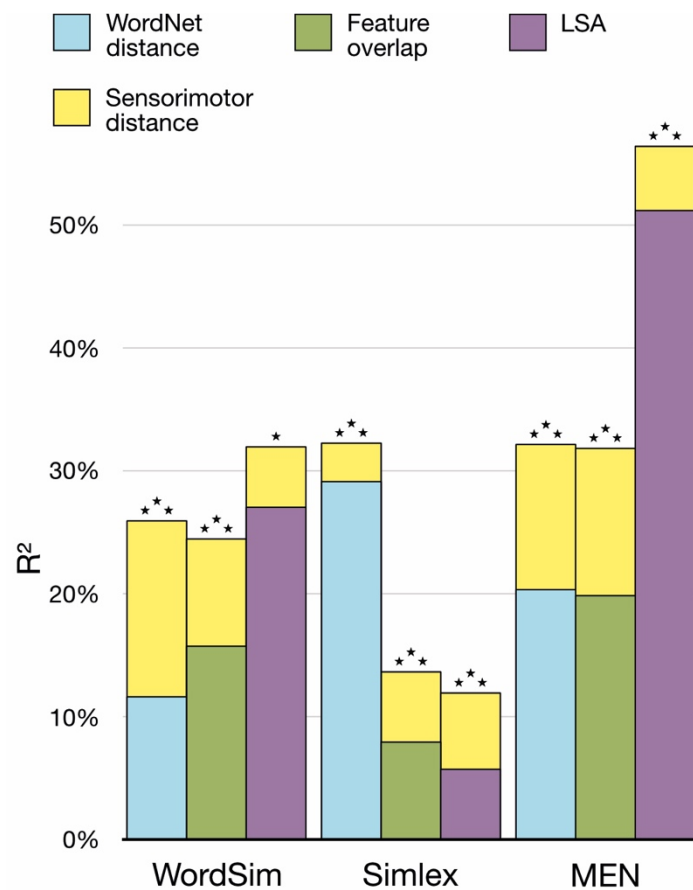


Figure B3. Unique effects of sensorimotor distance (top of stacked bar; yellow) in explaining variance in human similarity judgements when added to a regression model already containing an alternative measure of semantic similarity (bottom of stacked bar; color varies). Regressions were performed separately for each dataset (WordSim, Simlex, MEN) and alternative measure of semantic similarity (WordNet distance, feature overlap, LSA). Asterisks indicate evidence for including sensorimotor distance at Step 2 compared to alternative predictor at Step 1 (*: log BF₁₀ > log 10, **: log BF₁₀ > log 100, ***: log BF₁₀ > log 1000). (See Figure 6 for a comparison with CBOW.)













Dataset	Predictors				
	1 (Best)	2	3	4 (Weakest)	Not a predictor
WordSim	 LSA	 WordNet	 Sensorimotor distance	 Feature overlap	
Simlex	 WordNet	 Sensorimotor distance	 Feature overlap		 LSA
MEN	 LSA	 Sensorimotor distance	 Feature overlap	 WordNet	

Figure B4. Rank order from best to worst of each semantic similarity measure in predicting human similarity judgements across three datasets, based on inclusion Bayes Factors in best-model regressions. (See Figure 7 for a comparison with CBOW.)

Declarations

Availability of Data, Materials, and Code

All images, code, and data used in this article are available at <https://osf.io/d42q6/>, licensed under a Creative Commons Attribution 4.0 International License (CC-BY), which permits use, sharing, adaptation, distribution, and reproduction in any medium or format, so long as you give appropriate credit to the original authors and source, provide a link to the Creative Commons license, and indicate if changes were made. Sensorimotor distance measures are available via a web tool at <https://www.lancaster.ac.uk/psychology/smdistance/> under the same licence terms. To view a copy of the license, visit <http://creativecommons.org/licenses/by/4.0/>

Funding

This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 682848) to LC.

Conflicts of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Acknowledgements

We thank Erin Buchanan for permission to include feature overlap measures from Buchanan et al. (2019) in the datasets we make available.

References

- Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology*, *18*, 362–365. doi:10.1080/14640746608400055
- Banks, B., Wingfield, C., & Connell, L. (2021). Linguistic Distributional Knowledge and Sensorimotor Grounding both Contribute to Semantic Category Production. *Cognitive Science*, *45*(10). e13055. doi:10.1111/cogs.13055
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 577–660. doi:10.1017/S0140525X99002149
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media, Inc.
- Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Results*, *49*, 1–47. doi:10.1613/jair.4135
- Bruni, E. (2012, April 30). The MEN Test Collection [Online dataset]. Retrieved from <http://clic.cimec.unitn.it/~elia.bruni/MEN>.
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. doi:10.3758/s13428-013-0403-5
- Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, *51*(4), 1849–1863. doi:10.3758/s13428-019-01243-z
- Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, *44*(3), 890–907. doi:10.3758/s13428-011-0183-8

- Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8, 240–247. doi:10.1016/S0022-5371(69)80069-1
- Connell, L., Brand, J., Carney, J., Brysbaert, M., & Lynott, D. (2019). Go big and go grounded: Categorical structure emerges spontaneously from the latent structure of sensorimotor experience. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society* (p. 3434). Austin, TX: Cognitive Science Society.
- Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, 125, 452–465.
- Connell, L., & Lynott, D. (2014). Principles of representation: Why you can't represent the same concept twice. *Topics in Cognitive Science*, 6, 390-406.
doi:10.1016/j.cognition.2012.07.010
- Connell, L., Lynott, D., & Banks, B. (2018). Interoception: the forgotten modality in perceptual grounding of abstract and concrete concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(20170143), 1–9.
doi:10.1098/rstb.2017.0143
- Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132(2), 163–201. doi:10.1037/0096-3445.132.2.163
- Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, 41(8), 2149–2169. doi:10.1111/cogs.12453

- Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, 46(4), 1119–1127. doi:10.3758/s13428-013-0420-4
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems (TOIS)*, 20(1), 406–414. doi:10.1145/503104.503110
- Firth, J. R. (1957). *Studies in Linguistic Analysis*. Oxford, UK: Blackwell.
doi:10.2307/411592
- Gabrilovich, E. (2002, February 10) *The WordSimilarity-353 Test Collection* [Online dataset]. Retrieved from <http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/>
- Gagné, C. L., Spalding, T. L., Spicer, P., Wong, D., Rubio, B., & Cruz, K. P. (2020). Is buttercup a kind of cup? Hyponymy and semantic transparency in compound words. *Journal of Memory and Language*, 113, 104–110. doi:10.1016/j.jml.2020.104110
- Goldstone, R. L., & Son, J. Y. (2012). Similarity. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 155–176). Oxford University Press.
- Goodman, N. (1972). Seven Strictures on Similarity. In N. Goodman (Ed.), *Problems and Projects* (pp. 437–447). New York: Bobbs-Merrill.
- Günther, F., Petilli, M. A., Vergallito, A., & Marelli, M. (2020). Images of the unseen: Extrapolating visual representations for abstract and concrete words in a data-driven computational model. *Psychological Research*. doi:10.1007/s00426-020-01429-7
- Hahn, U. (2014). Similarity. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(3), 271–280. doi:10.1002/wcs.1282
- Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition*, 65(2–3), 137–165. doi:10.1016/S0010-0277(97)00042-5

- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346. doi:10.1016/0167-2789(90)90087-6
- Harris, Z. S. (1954). Distributional structure. *Word*, 10, 146–162.
doi:10.1080/00437956.1954.11659520
- Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41, 301–307.
[https://doi.org/10.1016/S0896-6273\(03\)00838-9](https://doi.org/10.1016/S0896-6273(03)00838-9)
- Hill, F. (n.d.). *SimLex-999* [Online dataset]. Retrieved from
<https://fh295.github.io/simlex.html>.
- Hill, F., Reichart, R., & Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41, 665–695.
doi:10.1162/COLI_a_00237
- Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E. J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3(2), 200–215. doi:10.1177/2515245919898657
- Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *Quarterly Journal of Experimental Psychology*, 61(7), 1036–1066. doi:10.1080/17470210701438111
- JASP Team (2020). JASP (Version 0.16) [Computer software].
- Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.
- Jolicoeur, P., Gluck, M. A., & Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology*, 16, 243–275. doi:10.1016/0010-0285(84)90009-4

- Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). *Psychological Review*, *122*(3), 570–574. doi:10.1037/a0039248
- Landauer, T. K., & Dumais, S. T. (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. doi:10.1037/0033-295X.104.2.211
- Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. *ArXiv Preprint: arXiv:1501.02598*.
- Lopopolo, A. & van Miltenburg, E. (2015). Sound-based distributional models. In Purver, M., Sadrzadeh, M., & Stone, M. (Eds.), *Proceedings of the 11th International Conference on Computational Semantics* (pp. 70–75). London, UK: Association for Computational Linguistics.
- Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302. doi:10.1111/j.1756-8765.2010.01106.x
- Louwerse, M. M., & Connell, L. (2011). A taste of words: Linguistic context and perceptual simulation predict the modality of words. *Cognitive Science*, *35*(2), 381–398. doi:10.1111/j.1551-6709.2010.01157.x
- Louwerse, M. M., & Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In M. de Vega, A. Glenberg, & A. C. Graesser (Eds.), *Symbols and Embodiment: Debates on Meaning and Cognition* (pp. 309–326). Oxford, UK: Oxford University Press.
- Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength

for 40,000 English words. *Behavior Research Methods*, 52, 1271–1291.

doi:10.3758/s13428-019-01316-z

Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, & Computers*, 36, 421–431. doi:10.3758/BF03195590

Mahalanobis, P. C. (1936). On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India, Volume 2* (pp. 49–55). Kolkata, India: National Institute of Science.

Mandera, P. (2016). *English, all words - CBOW model trained on a concatenation of UKWAC and subtitle corpus, 300 dimensions, window size 6* [Online dataset]. Retrieved from http://meshugga.ugent.be/snaut-downloads/spaces/english/predict/english-all.words-cbow-window.6-dimensions.300-ukwac_subtitle_en.w2v.gz

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, 57–78. doi:10.1016/j.jml.2016.04.001

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4), 547–559. doi:10.3758/BF03192726

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39–41. doi:10.1145/219717.219748

Miller, G. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

- Montefinese, M., Zannino, G. D., & Ambrosini, E. (2015). Semantic similarity between old and new items produces false alarms in recognition memory. *Psychological Research*, *79*, 785–794. doi:10.1007/s00426-014-0615-z
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.
- Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, UK: Oxford University Press.
- Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543). Stroudsburg, PA: Association for Computational Linguistics. doi:10.3115/v1/D14-1162
- Princeton University. (2010). About WordNet. *WordNet*. Princeton University.
- Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge: MIT Press.
- Raveh, M. (2002). The contribution of frequency and semantic similarity to morphological processing. *Brain and Language*, *81*(1–3), 312–325. doi:10.1006/brln.2001.2527
- Reilly, J., Flurie, M., & Peelle, J. E. (2020). The English lexicon mirrors functional brain activation for a sensory hierarchy dominated by vision and audition: Point-counterpoint. *Journal of Neurolinguistics*, *55*, 100895. <https://doi.org/10.1016/j.jneuroling.2020.100895>
- Ren, X., & Coutanche, M. N. (2021). Sleep reduces the semantic coherence of memory recall: An application of latent semantic analysis to investigate memory reconstruction. *Psychonomic Bulletin & Review*, *28*, 1336–1343. doi:10.3758/s13423-021-01919-8
- Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

- Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2), 303–345. doi:10.1111/j.1756-8765.2010.01111.x
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573–605. doi:10.1016/0010-0285(75)90024-9
- Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 100(5), 401–409. doi:0.1109/T-C.1969.222678
- Strube, M., & Ponzetto, S. P. (2006, July). WikiRelate! Computing semantic relatedness using Wikipedia. In *AAAI* (pp. 1419–1424). Boston, MA.
- van Hoef, R., Connell, L., & Lynott, D. (2019). The Role of Sensorimotor and Linguistic Information in the Basic-Level advantage. In A.K. Goel, C.M. Seifert, & C. reksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (p. 3376). Montreal, QB: Cognitive Science Society.
- Vinson, D.P., Vigliocco, G. (2008) Semantic feature production norms for a large set of objects and events. *Behavior Research Methods* 40, 183–190.
doi:10.3758/BRM.40.1.183
- Westbury, C. (2016). Pay no attention to that man behind the curtain: Explaining semantics without semantics. *The Mental Lexicon*, 11(3), 350–374. doi:10.1075/ml.11.3.02wes
- Wingfield, C., & Connell, L. (2022). Understanding the role of linguistic distributional knowledge in cognition. *Language, Cognition and Neuroscience*.
doi:10.1080/23273798.2022.2069278
- Wittgenstein, L. (1953). *Philosophical Investigations* (G. E. M. Anscombe, trans.). Macmillan Publishing Company.