English Semantic Feature Production Norms: An Extended Database of 4,436 Concepts

Erin M. Buchanan[1], K. D. Valentine[2], & Nicholas P. Maxwell[1]

[1] Missouri State University

[2] University of Missouri

Author Note

Erin M. Buchanan is an Associate Professor of Quantitative Psychology at Missouri State University. K. D. Valentine is a Ph.D. candidate at the University of Missouri. Nicholas P. Maxwell is a Masters' candidate at Missouri State University.

We would like to thank Keith Hutchison and David Balota for their contributions to this project, including the funds to secure Mechanical Turk participants. Additionally, we thank Simon De Deyne and an anonymous reviewer for their comments on this manuscript.

Correspondence concerning this article should be addressed to Erin M. Buchanan, 901 S. National Ave, Springfield, MO 65897. E-mail: erinbuchanan@missouristate.edu

Abstract

A limiting factor in understanding memory and language is often the availability of large numbers of stimuli to use and explore in experimental studies. In this study, we expand on three previous databases of concepts to over 4,000 words including nouns, verbs, adjectives, and other parts of speech. Participants in the study were asked to provide lists of features for each concept presented (a semantic feature production task), which were combined with previous research in this area. These feature lists for each concept were then coded into their root word form and affixes (i.e., *cat* and *s* for *cats*) to explore the impact of word form on semantic similarity measures, which are often calculated by comparing concept feature lists (feature overlap). All concept features, coding, and calculated similarity information is provided in a searchable database for easy access and utilization for future researchers when designing experiments that use word stimuli. The final database of word pairs was combined with the Semantic Priming Project to examine the relation of semantic similarity statistics on semantic priming in tandem with other psycholinguistic variables.

Keywords: semantics, word norms, database, psycholinguistics

English Semantic Feature Production Norms: An Extended Database of 4,436 Concepts

Semantic features are the focus of a large area of research which tries to delineate the semantic representation of a concept. These features are key to models of semantic memory (Collins & Loftus, 1975; i.e., memory for facts; Collins & Quillian, 1969), and they have been used to create both feature based (Cree & McRae, 2003; Smith, Shoben, & Rips, 1974; Vigliocco, Vinson, Lewis, & Garrett, 2004) and distributional based models (Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Riordan & Jones, 2011). Feature based models indicate that the degree of similarity between concepts is defined by their overlapping feature lists, while distributional based models posit that similarity is defined by the overlap between linguistic network or context. To create feature based similarity, participants were often asked to create lists of properties for categories of words. This property listing was a seminal task with corresponding norms that have been prevalent in the literature (Ashcraft, 1978; Rosch & Mervis, 1975; Toglia, 2009; Toglia & Battig, 1978). Feature production norms are created by soliciting participants to list properties or features of a target concept without focusing on category. These features are then compiled into feature sets that are thought to represent the memory representation of a particular concept, especially in early feature based models of memory (Collins & Loftus, 1975; Collins & Quillian, 1969; Jones, Willits, & Dennis, 2015; McRae & Jones, 2013).

For example, when queried on what features define a *cat*, participants may list *tail*, *animal*, and *pet*. These features capture the most common types of descriptions: “is a” and “has a”. Additionally, feature descriptions may include uses, locations, behavior, and gender (i.e., *actor* denotes both a person and gender). The goal of these norms is often to create a set of high-probability features, as there can and will be many idiosyncratic features listed in this task, to explore the nature of concept structure. In the classic view of category structure, concepts have defining features or properties, while the probabilistic view suggests that categories are fuzzy with features that are typical of a concept (Medin, 1989). These norms have now been published in Italian (Montefinese, Ambrosini, Fairfield, & Mammarella, 2013; Reverberi, Capitani, & Laiacona, 2004), German (and Italian, Kremer & Baroni, 2011), Portuguese (Stein & de Azevedo Gomes, 2009), Spanish (Vivas, Vivas, Comesaña, Coni, & Vorano, 2017), and Dutch (Ruts et al., 2004), as well as for the blind (Lenci, Baroni, Cazzolli, & Marotta, 2013).

Previous work on semantic feature production norms in English includes databases by McRae, Cree, Seidenberg, and McNorgan (2005), Vinson and Vigliocco (2008), Buchanan, Holmes, Teasley, and Hutchison (2013), and Devereux, Tyler, Geertzen, and Randall (2014). McRae et al. (2005)'s feature production norms focused on 541 nouns, specifically living and nonliving objects. Vinson and Vigliocco (2008) expanded the stimuli set by contributing norms for 456 concepts that included both nouns and verbs. Buchanan et al. (2013) broadened to concepts other than nouns and verbs with 1808 concepts normed. The Devereux et al. (2014) norms included a replication of McRae et al. (2005)'s concepts with the addition of several hundren more concrete concepts. The current paper represents over two thousand new concepts added to these previous projects and a reanalysis of the original data.

Creation of norms is vital to provide investigators with concepts that can be used in future research. The concepts presented in the feature production norming task are usually called *cues*, and the responses to the cue are called *features*. In a semantic priming task, the concept paired with a cue (first word) is denoted as a *target* (second word). In a lexical decision task, participants are shown cue words before a related or unrelated target word. Their task is to decide if the target word is a word or nonword as quickly as possible. A similar task, naming, involves reading the second target word aloud after viewing a related or unrelated cue word. Semantic priming occurs when the target word is recognized (responded to or read aloud) faster after the related cue word in comparison to the unrelated cue word (Moss et al., 1995). The feature list data created from the production task can be used to determine the strength of the relation between cue and target word, often by calculating the feature overlap, or number of shared features between concepts (McRae et al., 2005). Both the cue-feature lists and the cue-cue combinations (i.e., the relation between two cues in a feature production dataset, which becomes a cue-target combination in the priming task) are useful and important data for researchers in exploring various semantic based phenomena.

These feature category lists can provide insight into the probabilistic nature of language and conceptual structure (Cree & McRae, 2003; McRae, Sa, & Seidenberg, 1997; Moss, Tyler, & Devlin, 2002; Pexman, Holyk, & Monfils, 2003). Additionally, the feature production norms can be used as the underlying data to create models of semantic priming and cognition focusing on cue-target relation (Cree, McRae, & McNorgan, 1999; Rogers & McClelland, 2004; Vigliocco et al., 2004). When using database norms to select for stimuli, others have studied semantic word-picture interference (i.e., slower naming times when distractor words are related category concepts in a picture naming task; Vieth, Mcmahon, & Zubicaray, 2014), recognition memory (Montefinese, Zannino, & Ambrosini, 2015), and semantic richness, which is a measure of shared defining features (Grondin, Lupker, & McRae, 2009; Kounios et al., 2009; Yap & Pexman, 2016; Yap, Lim, & Pexman, 2015). The Vinson and Vigliocco labs have shown the power of turning in-house data projects into a larger norming set (Vinson & Vigliocco, 2008), as they published papers on aphasia (i.e., the loss of understanding speech; Vinson & Vigliocco, 2002; Vinson, Vigliocco, Cappa, & Siri, 2003), meaning-syntactic differences (Vigliocco, Vinson, & Siri, 2005; i.e., differences in naming times based on semantic or syntactic similarity; Vigliocco, Vinson, Damian, & Levelt, 2002), and representational models (Vigliocco et al., 2004).

However, it would be unwise to consider these norms as an exact representation of a concept in memory (McRae et al., 2005). These norms represent salient features that participants can recall, likely because saliency is considered special to our understanding of concepts (Cree & McRae, 2003). Additionally, Barsalou (2003) suggested that participants are likely creating a mental model of the concept based on experience and using that model to create a feature property list. This model may represent a specific instance of a category (i.e., their pet dog), and feature lists will represent that particular memory. One potential solution to overcome saliency effects would be to solicit applicability ratings for features across multiple exemplars (i.e., specific members) of a category, as De Deyne et al. (2008) have shown that this procedure provides reliable ratings across exemplars and provides more connections than the sparse representations that can occur when producing features.

Computational modeling of memory requires sufficiently large datasets to accurately portray semantic memory, therefore, the advantage of big data in psycholinguistics cannot be understated. There are many large corpora that could be used for exploring the structure of language and memory through frequency (see the SUBTLEX projects Brysbaert & New, 2009; New, Brysbaert, Veronis, & Pallier, 2007). Additionally, there are large lexicon projects that explore how the basic features of words affect semantic priming, such as orthographic neighborhood (words that are one letter different from the concept), length, and part of speech (Balota et al., 2007; Keuleers, Lacey, Rastle, & Brysbaert, 2012). In contrast to these basic linguistic features of words, other norming efforts have involved subjective ratings of concepts. Large databases of age of acquisition (i.e., rated age of learning the concept; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), concreteness (i.e., rating of how perceptible a concept is; Brysbaert, Warriner, & Kuperman, 2014), and valence (i.e., rating of emotion in a concept; Warriner, Kuperman, & Brysbaert, 2013) provide further avenues for understanding the impact these rated properties have on semantic memory. For example, age of acquisition and concreteness ratings have been shown to predict performance on recall tasks (Brysbaert et al., 2014; Dewhurst, Hitch, & Barry, 1998), while valence ratings are useful for gauging the effects of emotion on meaning (Warriner et al., 2013). These projects represent a small subset of the larger normed stimuli available (Buchanan, Valentine, & Maxwell, 2018), however, research is still limited by the overlap between these datasets. If a researcher wishes to control for lexical characteristics and subjective rating variables, the inclusion of each new variable to the study will further restrict the item pool for study. Large, overlapping datasets are crucial for exploring the entire range of an effect, and to ensure that the stimuli set is not the only contributing factor to the results of a study.

Therefore, the purpose of this study was to expand the number of cue and feature word stimuli available, which additionally increases the possible cue-target pairings for studies using word-pair stimuli (like semantic priming tasks). To accomplish these goals, we have expanded our original semantic feature production norms (Buchanan et al., 2013) to include all cues and targets from The Semantic Priming Project (Hutchison et al., 2013). The existing norms were reprocessed along with these new norms to provide new feature coding and affixes (i.e., word addition that modifies meaning, such as *pre* or *ing*) to explore the impact of word form. Previously, Buchanan et al. (2013) illustrated convergent validity with McRae et al. (2005) and Vinson and Vigliocco (2008) with a difference in approach to processing feature production data. In McRae et al. (2005) and Vinson and Vigliocco (2008), features were coded with complexity, matching the "is a" and "has a" format that was first found in Collins and Quillian (1969) and Collins and Loftus (1975) models. Buchanan et al. (2013) took a count based approach, wherein each feature is treated as a separate concept (i.e. *four legs* would be treated as two features, rather than one complex feature). Both approaches allow for the computation of similiarity by comparing feature lists for cue words, however, the count based approach matches popular computational models, such as Latent Semantic Analysis (Landauer & Dumais, 1997) and Hyperspace Analogue to Language (Lund & Burgess, 1996). These models treat each word in a document or text as a cue word and similarity is computed by assessing a matrix of frequency counts between concepts and texts, which is similar to comparing overlapping feature lists.

In the previous study, each feature was converted to common form if they denoted the same concept (i.e., most features were translated to their root form). This process often occurs to help capture the essential features without increasing the sparsity of the matrix (i.e., the matrix only contains one representation for *beauty*, rather than several for all word forms, thus lessening the number of empty cells in a cue-feature matrix). However, we previously included a few exceptions to this coding system, such as *act* and *actor* when the differences in features denoted a change of action (noun/verb) or gender or cue sets did not overlap (i.e., features like *will* and *willing* did not have overlapping associated cues). These exceptions were designed to capture how changes in morphology might be important cues to word meaning, as hybrid models of word identification have outlined that morpheme processing can be complex (Caramazza, Laudanna, & Romani, 1988; Marslen-Wilson, Tyler, Waksler, & Older, 1994). Hybrid models include both a compositional view (i.e., words are first broken down into their components *cat* and *s*; Jarvella & Meijers, 1983; MacKay, 1978) and a full-listing view (i.e., each word form is represented completely separately, *cat* and *cats* and processing occurs as a race between each type of representation, Bradley, 1980; Butterworth, 1983). Given these models and sparsity considerations, we created a coding system to capture the feature word meaning, in addition to morphology, to provide different levels of information about each cue-feature combination.

The entire dataset is available on our website (http://wordnorms.com/) which has been revamped with a new interface and web applications to easily find and select stimuli for future experiments. The data collection, (re)processing, website, and finalized dataset are detailed below. The basic properties of the cue-feature data will be detailed, such as the average number of features each cue elicited across parts of speech and datasets. The cue-feature data will be explored for divergent validity from the free association norms to show evidence that the new feature production norms provide additional information not found in the Nelson, McEvoy, and Schreiber (2004) dataset. We then provide details on how to calculate semantic similarity and then use these values to portray convergent validity by correlating multiple measures of semanticity. Additionally, the similarity measures are compared to the priming times from the Semantic Priming Project (Hutchison et al., 2013) to demonstrate the relation between semantic similarity and priming.

Method

Participants

Participants in the new stimuli set were recruited from Amazon's Mechanical Turk, which is a large, diverse participant pool wherein users can complete surveys for small sums of money (Buhrmester, Kwang, & Gosling, 2011). Answers can be screened for errors, and incorrect or incomplete surveys can be rejected or discarded without payment. Each participant was paid five cents for a survey, and they could complete multiple Human Intelligence Tasks or HITS. Participants were required to be located in the United States with a HIT approval rate of at least 80%, and no other special qualifications were required. HITS would remain active until $n = 30$ valid survey answers were obtained. Table 1 includes the sample sizes from the new study (Mechanical Turk 2), as well as the sample sizes from the previous study, as described in Buchanan et al. (2013).

Materials

A main purpose of this second norming set was to expand the Buchanan et al. (2013) norms to include all concepts from the Semantic Priming Project (Hutchison et al., 2013). The original concept set was selected primarily from the Nelson et al. (2004) database, with small overlaps in the McRae et al. (2005) and Vinson and Vigliocco (2008) database sets for convergent validity. In the Semantic Priming Project, cue-target pairs were shown to participants to examine naming (i.e., reading a concept aloud) and lexical decision (i.e., responding if a presented string is a word or nonword) response latency priming across related and unrelated pairs. The related pairs included first associate (most common response to a cue, *sum-add*) and other associates (second or greater common responses to cues, *safe-protect*) as their target words. The Buchanan et al. (2013) publication of concepts included many of the cue words from the Semantic Priming Project, while this project expanded to include unnormed cue words and all target words for all first and other associate pairs. The addition of these concepts allowed for complete overlap between the Semantic Priming Project and feature production norms. As mentioned earlier, the McRae et al. (2005) norms consist primarily of nouns, the Vinson and Vigliocco (2008) dataset includes nouns and verbs, while the Buchanan et al. (2013) included all word forms.

Concepts were labeled by part of speech using the English Lexicon Project (Balota et al., 2007), the free association norms, and Google's define search when necessary. When labeling these words, we used the most common part of speech to categorize concepts. This choice was predominately for simplicity of categorization, however, the participants were shown concepts without the suggestion of which sense to use for the word. Therefore, multiple senses (i.e., *bat* is noun and a verb) are embedded into the feature production norms, while the database is labeled with single parts of speech. The other parts of speech can be found in the English Lexicon Project or multiple other databases. This dataset was combined with McRae et al. (2005) and Vinson and Vigliocco (2008) feature production norms, which resulted in a combined total of 4437 concepts. 70.4% of concepts were nouns, 14.9% adjectives, 12.4% verbs, and 2.3% were other forms of speech, such as adverbs and conjunctions. The new concepts from this norming set only constituted: $n = 1916$ concepts, 72.0 nouns, 14.9% adjectives, 12.4% verbs, and 2.3% other parts of speech.

Procedure

Each HIT was kept to five concepts, and usual survey response times were between five to seven minutes. Each HIT was open until thirty participants had successfully completed the HIT and were paid the five cents for the HIT. HITS were usually rejected if they included copied definitions from Wikipedia, "I don't know", or the participant wrote a paragraph about the concept. These answers were discarded, as described below. The survey instructions were copied from McRae et al. (2005)'s Appendix B, which were also used in the previous publication of these norms. Because the McRae et al. (2005) data was collected on paper, we modified these instructions slightly. The original lines to write in responses were changed to an online text box response window. The detailed instructions additionally no longer contained information about how a participant should only consider the noun of the target concept, as the words in our study included multiple forms of speech and senses. Participants were encouraged to list the properties or features of each concept in the following areas: physical (looks, sounds, and feels), functional (uses), and categorical (belongings). The exact instructions were as follows:

*“We want to know how people read words for meaning. Please fill in features of the word that you can think of. Examples of different types of features would be: how it looks, sounds, smells, feels, or tastes; what it is made of; what it is used for; and where it comes from. Here is an example:

duck: is a bird, is an animal, waddles, flies, migrates, lays eggs, quacks, swims, has wings, has a beak, has webbed feet, has feathers, lives in ponds, lives in water, hunted by people, is edible

Complete this questionnaire reasonably quickly, but try to list at least a few properties for each word. Thank you very much for completing this questionnaire.“*

Participants signed up for the HITS through Amazon’s Mechanical Turk website and completed the study within the Mechanical Turk framework. Approved HITs were compensated through the Mechanical Turk system. All answers were then combined into a larger dataset.

Data Processing

The entire dataset, at each processing stage described here, can be found at: https://osf.io/cjyzw/. On our OSF page, we have included a detailed processing guide on how concepts were (re)examined for this publication. This paper was written with *R* markdown (R Core Team, 2017) and *papaja* (Aust & Barth, 2018). The markdown document allows an interested reader to view the scripts that created the article in line with the written text. However, the processing of the text documents was performed on the raw files, and therefore, we have included the processing guide for transparency of each stage.

First, each concept was separated into an individual text file that is included as the "raw" data online. Each of these files was then spell checked and corrected when the participant answer was obviously a typo. As noted earlier, participants often tried to cut and paste Wikipedia or other online dictionary sources into the their answers to complete surveys quickly with minimal effort. These entries were easily found because the formatting of the webpage was included in their answer. For example, the Wikipedia entry for *zerba* includes the phonetic spelling of the word, a set of paragraphs about zebras, a table of contents, and then sectioned paragraphs matching that table of contents. To find this data, lab members would open the raw text files that were compiled for each cue, look for these large blocks of formatted text, and delete that information. Approximately 113 HITS were rejected because of poor data, and 4524 HITS were paid. Therefore, we estimate approximately 2% of the HITS included Wikipedia articles or other ineligible entries. Next, each concept was processed for feature frequency. In this stage, the raw frequency counts of each cue-feature combination were calculated and put together into one large file. Cue-cue combinations were discarded, as participants might write "a zebra is a horse" when asked to define *zebra*. English stop words such as *the, an, of* were then discarded, as well as terms that were often used as part of a definition (*like, means, describes*).

We then created a “translated” column for each feature listed. This column indicated the root word for each feature, and additional columns were added with the affixes that were used in the original feature. For example, the original feature *cats* would be translated to *cat* and *s*, wherein *cat* would be the translated feature and the *s* would be the affix code. The translation was first started by using a Snowball type stemmer (Porter, 2001), written in Python by a colleague of the first author. All original features and their roots from this process were then put into an Excel document, which was reviewed by the first author for consistency and concepts with affixes that were not stemmed. Usually the noun version of the feature would be used for the translation or the most common part of speech for each feature.

At this stage, the data was reduced to cue-feature combinations that were listed by at least 16% of participants (matching McRae et al. (2005)’s procedure) or were in the top five features listed for that cue. This calculation was performed on the feature percent for the root word (the *translated* column). For example, *beauty* may have been listed as *beauty, beautiful, beautifully, beautifulness*, and this feature would have been listed four times in the dataset for the original cue (original feature in the *feature* column).

The sample size for the cue was added to this dataset, as the sample sizes varied across experiment time, as shown in Table 1. Therefore, instead of using raw feature frequency, we normalized each count into the percent of participants that included that feature with each cue. The *frequency_feature* column indicates the frequency of the original, unedited feature, while the *frequency_translated* includes all combinations of *beauty* into one overall feature. Because non-nouns can be more difficult to create a feature list for, we included the top five descriptors in addition to the 16% listed criteria, to ensure that each concept included at least five features. Table 2 indicates the average number of cue-feature pairs found for each data collection site/time point and part of speech for the cue word.

The parts of speech for the cue, original feature, and translated feature were merged with this file as described above. Table 3 depicts the pattern of feature responses for cue-feature part of speech combinations. This table includes the percent of features listed for each cue-feature part of speech combination (i.e., what is the percent of time that both the cue and feature were both adjectives) for the original feature (raw) and translated feature (root). Next, the average frequency percent was calculated along with their standard deviations. These columns indicate the percent that a cue-feature part of speech combination was listed across participants (i.e., what is the average percent of participants that listed an adjective feature for an adjective cue). These two types of calculation describe the likelihood of seeing part of speech combinations across the concepts, along with the likelihood of those cue-feature part of speech combinations across participants. Statistics in Table 3 only include information from the reprocessed Buchanan et al. (2013) norms and the new cues collected for this project.

The top cue-feature combinations for the reprocessed and new data collection were then combined with the cue-feature combinations from McRae et al. (2005) and Vinson and Vigliocco (2008). We included all the cue-feature combinations listed in their supplemental files with the feature in the raw feature column. If features could be translated into root words with affixes, the same procedure as described above was applied. The final file then included columns for the original dataset, cue, feature, translated feature, frequency of the original feature, frequency of the translated feature, sample size, and frequency percentages for the original and translated feature. The cue-feature file includes 69284 cue-raw feature combinations, where 48925 are from our dataset, and 24449 of which are cue-translated feature combinations.

The final data processing step was to code affixes found on the original features. Multiple affix codes were often needed for features, as *beautifully* would have been translated to *beauty*, *ful*, and *ly* (the *root, a1, and a2* columns; though, three affix columns were created in total). The research team searched lists of affixes online and collectively discussed how to code each affix, and the complete coding system can be found online in our OSF files. If an affix coding was unclear, the root and affix word were discussed in a lab meeting. Table 4 displays the list of affix types, common examples for each type of affix, and the percent of affixes that fell into each category. The percent values are calculated on the overall affix list, as feature words could have up to three different affixes. Generally, affixes were tagged in a one-to-one match, however, special care was taken with numbers and verb tenses, and the lead author checked these categories after lab member coding. Features like cat*s* would be coded as a number affix, while features like walk*s* would be coded as a third person verb.

In the final words file found online, we additionally added forward strength (FSG) and backward strength (BSG) for investigation into association overlap (Nelson et al., 2004). Forward strength indicates the number of times a target word was listed in response to a cue word in a free association task, which simply asks participants to name the first word that comes to mind when presented with a cue word. Backward strength is the number of times a cue word was listed with a target word, as free association is directional (i.e., the number of times *cheese* is listed in response to *cheddar* is not the same as the number of times that *cheddar* is listed in response to *cheese*). The last few columns indicate the word list a concept was originally normed in to allow for matching to the original raw files on the OSF page, along with the code for each school and time point of collection.

Both forms of the feature are provided for flexibility in calculating overlap by using the original feature (raw), the translated feature (root), and the affix overlap by code (affix). Cosine values were calculated for each of these feature sets by using the following formula:

$$\frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

This formula is similar to a dot-product correlation, where A_i and B_i indicate the overlapping feature frequency (normalized, therefore, the percent) between cue A and cue B.

The *i* subscript denotes the current cue, and when features match, the frequencies are multiplied together and summed across all matches (Σ). For the denominator, the feature frequency is first squared and summed from *i* to *n* features for cue A and B. The square root of these summation values is then multiplied together. In essence, the numerator calculates the overlap of feature frequency for matching features, while the denominator accounts for the entire feature frequency set for each cue. Cosine values range from 0 (no overlapping features) to 1 (complete overlapping features). With over four thousand cue words, just under twenty million cue-cue cosine combinations can be calculated. In the datasets presented online, we only included cue-cue combinations with a feature overlap of at least two features, in order to reduce the large quantity of zero and very low cosine values. This procedure additionally allowed for online presentation of the data, as millions of cosines were not feasible for our server. The complete feature list, along with our code to calculate cosine, can be used to obtain values not presented in our data if desired.

Website

In addition to our OSF page, we present a revamped website for this data at http://www.wordnorms.com/. The single word norms page includes information about each of the cue words including cue set size, concreteness, word frequency from multiple sources, length, full part of speech, orthographic/phonographic neighborhood, and number of phonemes, syllables, and morphemes. These values were taken from Nelson et al. (2004), Balota et al. (2007), and Brysbaert and New (2009). A definition of each of these variables is provided along with the minimum, maximum, mean, and standard deviation of numeric values. The table is programmed using Shiny apps (Chang, Cheng, Allaire, Xie, & McPherson, 2017). Shiny is an *R* package that allows the creation of dynamic graphical user interfaces for interactive web applications. The advantage to using Shiny applications is data manipulation and visualization with the additional bonus of up to date statistics for provided data (i.e., as typos are fixed or data is updated, the web app will display the most recent calculations). In addition to the variable table, users can search and save filtered output using our Shiny search app. With this app, you can filter for specific variable ranges and save the output in a csv or Excel file. The complete data is also provided for download.

On the word pair norms page, all information about word-pair statistics can be found. A second variable table is provided with semantic and associative statistics. This dataset includes the cue and target words from this project (cue-cue combinations), the root, raw, and affix cosines described above, as well as the original Buchanan et al. (2013) cosines. Additional semantic information includes Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) and JCN (JCN stands for Jiang-Conrath, see explanation below; Jiang & Conrath, 1997) values provided in the Maki, McKinley, and Thompson (2004) norms, along with forward strength and backward strength (FSG; BSG) from the Nelson et al. (2004) norms for association. The definitions, minimum, maximum, mean, and standard deviations of these values are provided in the app. Again, the searchable app includes all of these stimuli for cue-cue combinations with two or more features in common, where you can filter this data for experimental stimuli creation. The separation of single and word-pair data (as well as cosine calculation reduction to cues with two features in common) was practical, as the applications run slowly as a factor of the number of rows and columns of data. On each page, we link the data, applications, and source code so that others may use and manipulate our work depending on their data creation or visualization goals.

Results

An examination of the results of the cue-feature lists indicated that the new data collected was similar to the previous semantic feature production norms. As shown in Table 2, the new Mechanical Turk data showed roughly the same number of listed features for each cue concept, usually between five to seven features. These numbers represent, for each cue and part of speech, the average number of distinct cue-feature pairs provided by participants after processing. Table 3 portrayed that adjective cues generally included other adjectives or nouns as features, while noun cues were predominately described by other nouns. Verb cues included a large feature list of nouns and other verbs, followed by adjectives and other word forms. Lastly, the other cue types generally elicited nouns and verbs. Frequency percentages were generally between seven and twenty percent when examining the raw words. These words included multiple forms, as the percent increased to around thirty percent when features were translated into their root words. Indeed, nearly half of the 48925 cue-feature pairs were repeated, as 24449 cue-feature pairs were unique when examining translated features. Generally, because of the translation process, word forms shifted towards nouns and verbs and away from adjectives because adjectives are often formed by adding an affix to a noun or verb.

36030 affix values were found, which arose from 4407 of the 4436 cue concepts. 33052 first affixes were found, with 2832 second place affixes, and 146 third place affixes. Table 4 shows the distribution of these affix values. Generally, numbers were the largest category of affixes demonstrating that participants often indicated the quantity of the feature when describing the cue word. The second largest affix category was characteristics which denoted the switch to or from a noun form of the feature word (i.e., *angry* to *anger*). Verb tenses (past tense, present participle, and third person) comprised a large set of affixes indicating the type of concept or when a concept might be doing an action for a cue. Persons and objects affixes were used about 7% of the time on features to explain cues, while actions and processes were added to the feature about 8% of the time.

Divergent Validity

When collecting semantic feature production norms, there can be a concern that the information produced will simply mimic the free association norms, and thus, be a more of representation of association (context) rather than semanticity (meaning). Association and semanticity do overlap, however, the variables used to represent these concepts have been shown to tap different underlying constructs (Maki & Buchanan, 2008). Therefore, it is important to show that, while some overlap is expected, the semantic feature production norms provide useful, separate information from the free association norms. Table 5 portrays the overlap with the Nelson et al. (2004) norms. The percent of time a cue-feature combination was present in the free association norms was calculated, along with the average forward strength for those overlapping pairs. First, these values were calculated on the complete dataset with the McRae et al. (2005) and Vinson and Vigliocco (2008) norms–as we are presenting them as a combined dataset–on the translated cue-feature set only. Because we used the translated cue-feature set, repeated instances of cue-features would occur (i.e., the original *abandon-leave* and *abandon-leaving* becomes two lines when only using translated *abandon-leave*), and thus only the unique set was considered. Second, we calculated these values on each dataset separately, as well as for the 26 cues that overlapped in all three datasets.

The overall overlap between the database cue-feature sets and the free association cue-target sets was approximately 37%, ranging from 32% for verbs and nearly 52% for adjectives. Similar to our previous results, the range of the forward strength was large (.01 - .94), however, the average forward strength was low for overlapping pairs, M = .11 (SD = .14). These results indicated that while it will always be difficult to separate association and meaning, the dataset presented here represents a low association when examining overlapping values, and more than 60% of the data is completely separate from the free association norms. The limitation to this finding is the removal of idiosyncratic responses from the Nelson et al. (2004) norms, but even if these were to be included in some form, the average forward strength would still be quite low when comparing cue-feature lists to cue-target lists. In examining these values by dataset, it appears that the new norms have the highest overlap with the Nelson et al. (2004) data, while the average, standard deviation, minimum, and maximum values were roughly similar for each dataset and the overlapping cues. This effect is likely driven by the inclusion of adjectives and other forms of speech, which show higher overlaps than nouns and verbs, which represent the cues present in McRae et al. (2005) and Vinson and Vigliocco (2008).

In the last column of Table 5, we calculated the correlation between forward strength and the frequency percent for the the root (translated) cue-feature pairs. This correlation provides information the relation between the strength of the association and the frequency of cue-feature mentions. Correlations were similar across parts of speech except, notably, the other category included the lowest relation. This result is likely because the instructions of a semantic feature production task might exclude normal “first word that pops into your mind” association task concepts. The correlations across datasets and the overlapping cues were also similar, denoting that as forward strength increased, the likelihood of the cue-feature mentions also increased. In general, these cue-feature pairs were still of low associative strength, as shown in the mean column of Table 5.

Convergent Validity

To examine the validity of cosine values, we calculated the average cosine score between the new processing of the data for each of the three feature production norms used in this project. Overlapping cues in all of the three databases were found (n = 188), and the average cosine between their feature sets was examined. Buchanan et al. (2013) and the new dataset are listed with the subscript B, while McRae et al. (2005) is referred to with M and V for Vinson and Vigliocco (2008). For root cosine values, we found high overlap between all three datasets: M_{BM} = .67 (*SD* = .14), M_{BV} = .66 (*SD* = .18), and M_{MV} = .72 (*SD* = .11). The raw cosine values also correlated, even though the McRae et al. (2005) and Vinson and Vigliocco (2008) datasets were already mostly preprocessed for word stems: M_{BM} = .55 (*SD* = .15), M_{BV} = .54 (*SD* = .20), and M_{MV} = .45 (*SD* = .19). Last, the affix cosines overlapped similarly between Buchanan et al. (2013) and McRae et al. (2005) datasets, $M_{BM} = .43$ ($SD = .29$), but did not overlap with the Vinson and Vigliocco (2008) datasets: $M_{BV} = .04$ ($SD = .14$), and $M_{MV} = .09$ ($SD = .19$), likely due to Vinson and Vigliocco (2008) dataset preprocessing.

The correlation between root, raw, affix, previously found cosine, Latent Semantic Analysis score (LSA), and Jiang-Conrath semantic distance (JCN) were calculated to examine convergent validity. LSA is one of the most well-known semantic memory models (Landauer & Dumais, 1997; McRae & Jones, 2013), wherein a large text corpus (i.e., many texts) is used to create a word by document (i.e., each text) matrix. From this matrix, words are weighted relative to their frequency, and singular value decomposition is then used to select only the largest semantic components. This process creates a word space that can then be used to calculate the relation between two cues by examining the patterns of their occurrence across documents, usually cosine or correlation. JCN is calculated from an online dictionary (WordNet; Fellbaum & Felbaum, 1998), by measuring the semantic distance between concepts in a hierarchical structure. JCN is backwards coded, as zero values indicate close semantic neighbors (low dictionary distance) and high values indicate low semantic relation. These two measures were selected for convergent validity because they are well-cited measures of semanticity. To examine if the type of processing impacted convergent validity of the dataset, we calculcated the McRae et al. (2005) and Vinson and Vigliocco (2008) cosine values based on their original cue-feature matrices provided in their publications. These datasets were coded for more complex features in a propositional style ("is a", "has a"), while our processing took a single word count based approach. Therefore, providing the original processing correlations allows one to examine if the cosine values provided are covergent, as well as similarly correlated across other measures of semanticity.

As shown in Table 6, the intercorrelations between the cosine measures (root, raw, affix) are high, especially between our previous work and this dataset. We found that the correlation between processing styles was high and matched the intercorrelations between the new cosine measures (indicating convergent validity of coding style). The small negative correlations between JCN and cosine measures replicated previous findings (Buchanan et al., 2013). LSA values showed small positive correlations with cosine values, indicating some overlap with thematic information and semantic feature overlap (Maki & Buchanan, 2008). These correlations were slightly different than our previous publication, likely because here we restricted this cosine set to values with at least two features in common. LSA and JCN correlations were lower than LSA-cosine and JCN-cosine, but these values indicated that themes and dictionary distance were similarly related to feature overlap. Last, the correlation between propositional processing ("MV COS" column) and JCN was higher than the new root cosine measure (-.39 versus -.18 respectively). JCN is created through a hierarchical dictionary with a structure similar to the complex propositional coding provided in McRae et al. (2005) and Vinson and Vigliocco (2008), and correspondingly, the relation between them is stronger.

Relation to Semantic Priming

As a second examination of convergent validity, the correlation between values calculated from these norms and the Z-priming values from the Semantic Priming Project were examined. The Semantic Priming Project includes lexical decision and naming response latencies for priming at 200 and 1200 ms stimulus onset asynchronies (SOA). In these experiments, participants were shown cue-target words that were either the first associate of a concept or an other associate (second response or higher in the Nelson et al. (2004) norms) with the delay between the cue and target matching either 200 or 1200 ms (SOA). The response latency of the target word in the related condition (either first or other associate) was subtracted from the response latency in the unrelated condition to create a priming response latency. Therefore, each target item received four (two SOAs by two tasks: lexical decision or naming) priming times. We selected the Z-scored priming from the dataset to correlate with our data, as Hutchison et al. (2013) demonstrated that the Z-scored data more accurately captures priming controlled for individual differences in reaction times.

In addition to root, raw, and affix cosine, we additionally calculated feature set size for the cue and target of the primed pairs. Feature set size is the number of features listed by participants when creating the norms for that concept. Because of the nature of our norms, we calculated both feature set size for the raw, untranslated features, as well as the translated features. The average feature set sizes for our dataset can be found in Table 2. The last variable included was cosine set size which was defined as the number of other concepts each cue or target was nonzero paired with in the cosine values. Feature set size indicates the number of features listed for each cue or target, while cosine set size indicates the number of other semantically related concepts for each cue or target. Feature and cue set size are often called semantic richness, representing the variability or extent of associated information for a cue (Buchanan, Westbury, & Burgess, 2001; Pexman, Hargreaves, Edwards, Henry, & Goodyear, 2007; Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008). Several studies have showed the positive effects of semantic richness on semantic tasks based on task demand (Duñabeitia, Avilés, & Carreiras, 2008; Pexman et al., 2008; Yap, Pexman, Wellsby, Hargreaves, & Huff, 2012; Yap, Tan, Pexman, & Hargreaves, 2011), and thus, they were included as important variables to examine.

Tables 7 (for the lexical decision task) and 8 (for the naming task) display the correlations between the new semantic variables described above, as well as forward strength, backward strength, Latent Semantic Analysis score, and Jiang-Conrath semantic distance for reference. Only cue-target pairs with complete values were included in this analysis to allow for comparison between correlations. For lexical decision priming, we found small correlations between the root and raw cosine values and priming, with the largest for first associates in the 200 ms condition. The correlations decreased for the 1200 ms condition and the other associate SOAs. These two variables are highly correlated, therefore, it is not surprising that they have similar correlations with priming. Affix cosine also was slightly related to priming, especially for first associates in the 200 ms condition. Most of the cue and feature set sizes were not related to priming, showing correlations close to zero in most instances. Cue set size for the cue word was somewhat related to 200 ms priming, along with raw cue feature set size (for first associates only). These correlations are small, but they are comparable or greater than the correlations for association and other measures of semantic or thematic relatedness. For naming, the results are less consistent. Cosine values are related to 1200 ms naming in first associates, but none of the feature or cue set sizes showed any relationship with priming. Again, we see that many of the other associative and semantic variables correspondingly do not correlate with priming. In both naming and lexical decision priming, backward strength has a small but consistent relationship with priming, which may indicate the processing of the target back to the cue. Latent Semantic Analysis score was also a small predictor of priming across conditions.

As mentioned in the Website section, we have provided the data to calculate a broad range of information of linguistic information or simply use the provided values. From our OSF page (also linked to GitHub: https://github.com/doomlab/Word-Norms-2), you can find the data at each stage of processing and final data from this manuscript. Interested researchers could use our raw feature files to create their own coding schemes (or ones similar to McRae et al. (2005)), use the processed files to calculate set sizes for each cue or feature, and use these files plus the cosine files to create their own experimental stimuli (also avaliable as a Shiny app on http://www.wordnorms.com). These data could also be used to calculate other measures of interest, such as pointwise positive mutual information, entropy, and random walk statistics (De Deyne, Navarro, Perfors, & Storms, 2016).

Discussion

This research project focused on expanding the availability of English semantic feature overlap norms, in an effort to provide more coverage of concepts that occur in other large database projects like the Semantic Priming and English Lexicon Projects. The number and breadth of linguistic variables and normed databases has increased over the years, however, researchers can still be limited by the concept overlap between them. Projects like the Small World of Words provide newly expanded datasets for association norms, and our work helps fill the voids for corresponding semantic norms. To provide the largest dataset of similar data, we combined the newly collected data with previous work by using Buchanan et al. (2013), McRae et al. (2005), and Vinson and Vigliocco (2008) together. These norms were reprocessed from previous work to explore the impact of feature coding for feature overlap. As shown in the correlation between root and raw cosines, the parsing of words to root form created very similar results across other variables. This finding does not imply that these cosine values are the same, as root cosines were larger than their corresponding raw cosine. It does, however, imply that the cue-feature coding can produce similar results in raw or translated format. Because the correlation between the current paper's cosine values and the previous cosine values was nearly 1, we would suggest using the new values, simply for the increase in dataset size.

Of particular interest was the information that is often lost when translating raw features back to a root word. One surprising result in this study was the sheer number of affixes present on each cue word. With these values, we believe we have captured some of the nuance that is often discarded in this type of research. Affix cosines were less related to their feature root and raw counterparts, but also showed small correlations with semantic priming. Potentially, affix overlap can be used to add small, but meaningful predictive value to related semantic phenomena. Further investigation into the compound prediction of these variables is warranted to fully explore how these, and other lexical variables, may be used to understand semantic priming. An examination of the cosine values from the Semantic Priming Project cue-target set indicates that these values were low, with many zeros (i.e., no feature overlap between cues and targets). This restriction of range of the cosine relatedness could explain the small correlations with priming because the semantic priming was variable, but the cosine values were not.

One important limitation of the instructions in this study is that multiple senses of concepts were not distinguished. We did not wish to prime participants for specific senses to capture the features for multiple senses of a concept, however, this procedure could lead to lower cosine values for concepts that might intuitively seem very related. The feature production lists could be used to sort senses and recalculate overlap values, but it is likely that feature information is correspondingly mixed or sorted into small sublists in memory as well. The addition of the coded affix information may help capture some of those sense differences, as well as some of the spatial and relational features that are not traditionally captured by simple feature production. For example, by understanding the numbers or actors affixes, we may gain more information about semanticity that is often regarded as something to disregard in data processing.

We encourage readers to use the corresponding website associated with these norms to download the data, explore the Shiny apps, and use the options provided for controlled experimental stimuli creation. We previously documented the limitations of feature production norms that rely on on single word instances as their features (i.e., *four* and *legs*), rather than combined phrase sets. One potential limitation, then, is the inability to create fine distinctions between cues; however, the small feature set sizes imply that the granulation of features is large, since many distinguishing features are often never listed in these tasks. For instance, *dogs* are living creatures, but *has lungs* or *has skin* would usually not be listed during a feature production task, and thus, feature sets should not be considered a complete snapshot of mental representation (Rogers & McClelland, 2004). Additionally, the cue-feature lists could be explored for the type of cue-feature representation that is listed for each part of speech (i.e., physical, functional, etc.) and the complexity in coding could be increased or decreased depending on researcher goal. The previous data and other norms were purposely combined in the recoded format, so that researchers could use the entire set of available norms which increases comparability across datasets. Given the strong correlation between databases, we suspect that using single word features does not reduce their reliability and validity. We found high correlations between the different types of feature coding (i.e., complex/propositional versus single word/count), thus suggesting that either dataset could be used for future work where the advantage of the current project is the size of the norms.

References

Ashcraft, M. H. (1978). Property norms for typical and atypical items from 17 categories: A description and discussion. *Memory & Cognition*, *6*(3), 227–232. doi:10.3758/BF03197450

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. doi:10.3758/BF03193014

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *358*(1435), 1177–1187. doi:10.1098/rstb.2003.1319

Bradley, D. (1980). Lexical representation of derivational relation. In M. Aronoff & M. L. Kean (Eds.), *Juncture* (pp. 37–55). Saratoga, CA: Anma Libri.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. doi:10.3758/BRM.41.4.977

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., Holmes, J. L., Teasley, M. L., & Hutchison, K. A. (2013). English semantic word-pair norms and a searchable Web portal for experimental stimulus creation. *Behavior Research Methods*, *45*(3), 746–757. doi:10.3758/s13428-012-0284-z

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2018). LAB: Linguistic Annotated Bibliograpy - A searchable portal for normed database information. Retrieved from https://osf.io/r6y3n

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*, 531–544.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk. *Perspectives on Psychological Science*, *6*(1), 3–5. doi:10.1177/1745691610393980

Butterworth, B. (1983). Lexical representation. In B. Butterworth (Ed.), *Language production, vol. ii: Development, writing and other language processes* (pp. 257–294). London: Academic.

Caramazza, A., Laudanna, A., & Romani, C. (1988). Lexical access and inflectional morphology. *Cognition*, *28*(3), 297–332. doi:10.1016/0010-0277(88)90017-0

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2017). *Shiny: Web application framework for r*. Retrieved from https://CRAN.R-project.org/package=shiny

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. doi:10.1037/0033-295X.82.6.407

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247. doi:10.1016/S0022-5371(69)80069-1

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

Cree, G. S., McRae, K., & McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, *23*, 371–414. doi:10.1016/S0364-0213(99)00005-1

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General*, *145*(9), 1228–1254. doi:10.1037/xge0000192

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048. doi:10.3758/BRM.40.4.1030

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech , Language and the Brain (CSLB) concept property norms, 1119–1127. doi:10.3758/s13428-013-0420-4

Dewhurst, S. A., Hitch, G. J., & Barry, C. (1998). Separate effects of word frequency and age of acquisition in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(2), 284–298. doi:10.1037/0278-7393.24.2.284

Duñabeitia, J. A., Avilés, A., & Carreiras, M. (2008). NoA's ark: Influence of the number of associates in visual word recognition. *Psychonomic Bulletin & Review*, *15*, 1072–1077.

Fellbaum, C., & Felbaum, C. (1998). *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation.

Psychological Review, *114*(2), 211–244. doi:10.1037/0033-295X.114.2.211

Grondin, R., Lupker, S. J., & McRae, K. (2009). Shared features dominate semantic richness effects for concrete concepts. *Journal of Memory and Language*, *60*(1), 1–19. doi:10.1016/j.jml.2008.09.001

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., … Buchanan, E. M. (2013). The semantic priming project. *Behavior Research Methods*, *45*(4), 1099–1114. doi:10.3758/s13428-012-0304-z

Jarvella, R., & Meijers, G. (1983). Recognizing morphemes in spoken words: Some evidence for a stem-organized mental lexicon. In G. B. Flores d'Arcaos & R. Jarvella (Eds.), *The process of language understanding* (pp. 81–112). New York: Wiley.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*. Retrieved from http://arxiv.org/abs/cmp-lg/9709008

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37. doi:10.1037/0033-295X.114.1.1

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of Semantic Memory. *Oxford Handbook of Mathematical and Computational Psychology*, 232–254. Retrieved from http://psycnet.apa.org/psycinfo/2004-17297-001

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. doi:10.3758/s13428-011-0118-4

Kounios, J., Green, D. L., Payne, L., Fleck, J. I., Grondin, R., & McRae, K. (2009).

Semantic richness and the activation of concepts in semantic memory: Evidence from event-related potentials. *Brain Research*, *1282*, 95–102. doi:10.1016/j.brainres.2009.05.092

Kremer, G., & Baroni, M. (2011). A set of semantic norms for German and Italian. *Behavior Research Methods*, *43*(1), 97–109. doi:10.3758/s13428-010-0028-x

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. doi:10.3758/s13428-012-0210-4

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

Lenci, A., Baroni, M., Cazzolli, G., & Marotta, G. (2013). BLIND: A set of semantic feature norms from the congenitally blind. *Behavior Research Methods*, *45*(4), 1218–1233. doi:10.3758/s13428-013-0323-4

Lund, K., & Burgess, C. (1996). Hyperspace analogue to language (HAL): A general model semantic representation. *Brain and Cognition*, *30*(3), 5–5.

MacKay, D. G. (1978). Derivational rules and the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *17*, 61–71.

Maki, W. S., & Buchanan, E. M. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review*, *15*(3), 598–603. doi:10.3758/PBR.15.3.598

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*(3), 421–431. doi:10.3758/BF03195590

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, *101*(1), 3–33. doi:10.1037/0033-295X.101.1.3

McRae, K., & Jones, M. (2013). Semantic Memory. In D. Reisberg (Ed.),. Oxford University Press. doi:10.1093/oxfordhb/9780195376746.013.0014

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

McRae, K., Sa, V. R. de, & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, *126*(2), 99–130. doi:10.1037/0096-3445.126.2.99

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, *44*(12), 1469–1481. doi:10.1037/0003-066X.44.12.1469

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2013). Semantic memory: A feature-based analysis and new norms for Italian. *Behavior Research Methods*, *45*(2), 440–461. doi:10.3758/s13428-012-0263-4

Montefinese, M., Zannino, G. D., & Ambrosini, E. (2015). Semantic similarity between old and new items produces false alarms in recognition memory. *Psychological Research*, *79*(5), 785–794. doi:10.1007/s00426-014-0615-z

Moss, H. E. H., Ostrin, R. K. R., Tyler, I., Marlsen-Wilson, W., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning,* *Memory, and Cognition*, *21*(4), 863–883. doi:10.1037/ 0278-7393.21.4.863

Moss, H. E., Tyler, L. K., & Devlin, J. T. (2002). The emergence of category-specific deficits in a distribuited semantic system. In E. Forde & G. Humphreys (Eds.), *Category-specificity in mind and brain* (pp. 115–145). CRC Press.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. doi:10.3758/BF03195588

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661–677. doi:10.1017/S014271640707035X

Pexman, P. M., Hargreaves, I. S., Edwards, J. D., Henry, L. C., & Goodyear, B. G. (2007). The Neural Consequences of Semantic Richness When More Comes to Mind , Less Activation Is Observed. *Psychological Science*, *18*(5), 401–406.

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich : Effects of three measures of semantic, *15*(1), 161–167. doi:10.3758/PBR.15.1.161

Pexman, P. M., Holyk, G. G., & Monfils, M.-H. (2003). Number-of-features effects and semantic processing. *Memory & Cognition*, *31*(6), 842–855. doi:10.3758/BF03196439

Porter, M. (2001). Snowball: A language for stemming algorithms - Snowball. Retrieved from https://snowballstem.org/texts/introduction.html

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reverberi, C., Capitani, E., & Laiacona, E. (2004). Variabili semantico lessicali relative a tutti gli elementi di una categoria semantica: Indagine su soggetti normali italiani per la categoria “frutta". *Giornale Italiano Di Psicologia*, *31*, 497–522.

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

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* MIT Press.

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

Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T., & Storms, G. (2004). Dutch norm data for 13 semantic categories and 338 exemplars. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 506–515. doi:10.3758/BF03195597

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, *81*(3), 214–241. doi:10.1037/h0036351

Stein, L., & de Azevedo Gomes, C. (2009). Normas Brasileiras para listas de palavras associadas: Associação semântica, concretude, frequência e emocionalidade. *Psicologia: Teoria E Pesquisa*, *25*, 537–546. doi:10.1590/S0102-37722009000400009

Toglia, M. P. (2009). Withstanding the test of time: The 1978 semantic word norms. *Behavior Research Methods*, *41*(2), 531–533. doi:10.3758/BRM.41.2.531

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms.* Hillside, NJ:

Earlbaum.

Vieth, H. E., Mcmahon, K. L., & Zubicaray, G. I. D. (2014). The roles of shared vs . distinctive conceptual features in lexical access, *5*(September), 1–12. doi:10.3389/fpsyg.2014.01014

Vigliocco, G., Vinson, D. P., & Siri, S. (2005). Semantic and grammatical class effects in naming actions. *Cognition*, *94*, 91–100. doi:10.1016/j.cognition.2004.06.004

Vigliocco, G., Vinson, D. P., Damian, M. M. F., & Levelt, W. (2002). Semantic distance effects on object and action naming. *Cognition*, *85*, 61–69. doi:10.1016/S0010-0277(02)00107-5

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. F. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology*, *48*(4), 422–488. doi:10.1016/j.cogpsych.2003.09.001

Vinson, D. P., & Vigliocco, G. (2002). A semantic analysis of noun-verb dissociations in aphasia. *Journal of Neurolinguistics*, *15*, 317–351. doi:10.1016/S0911-6044(01)00037-9

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*(1), 183–190. doi:10.3758/BRM.40.1.183

Vinson, D. P., Vigliocco, G., Cappa, S., & Siri, S. (2003). The breakdown of semantic knowledge: Insights from a statistical model of meaning representation. *Brain and Language*, *86*(3), 347–365. doi:10.1016/S0093-934X(03)00144-5

Vivas, J., Vivas, L., Comesaña, A., Coni, A. G., & Vorano, A. (2017). Spanish semantic feature production norms for 400 concrete concepts. *Behavior Research Methods*, *49*(3), 1095–1106. doi:10.3758/s13428-016-0777-2

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. doi:10.3758/s13428-012-0314-x

Yap, M. J., & Pexman, P. M. (2016). Semantic richness effects in syntactic classification: The role of feedback. *Frontiers in Psychology*, *7*(July), 1394. doi:10.3389/fpsyg.2016.01394

Yap, M. J., Lim, G. Y., & Pexman, P. M. (2015). Semantic richness effects in lexical decision: The role of feedback. *Memory & Cognition*, *43*(8), 1148–1167. doi:10.3758/s13421-015-0536-0

Yap, M. J., Pexman, P. M., Wellsby, M., Hargreaves, I. S., & Huff, M. J. (2012). An abundance of riches : cross-task comparisons of semantic richness effects in visual word recognition. *Frontiers in Human Neuroscience*, *6*, 1–10. doi:10.3389/fnhum.2012.00072

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin and Review*, *18*(4), 742–750. doi:10.3758/s13423-011-0092-y

Table 1

Sample Size and Concept Norming Size for Each Data Collection Location/Time Point

Institution	Total Participants	Concepts	Mean *N*
University of Mississippi	749	658	67.8
Missouri State University	1420	720	71.4
Montana State University	127	120	63.5
Mechanical Turk 1	571	310	60
Mechanical Turk 2	198	1914	30

Table 2

Average (SD) Cue-Feature Pairs by Location/Time Point

Institution	Adjective	Noun	Verb	Other	Total
University of Mississippi	5.57 (1.53)	7.35 (4.05)	5.33 (0.87)	6.01 (2.11)	6.71 (3.44)
Missouri State University	5.74 (1.56)	6.85 (2.82)	6.67 (2.08)	7.45 (5.35)	6.65 (2.92)
Montana State University	5.81 (1.74)	7.25 (3.35)	5.59 (1.13)	5.76 (1.74)	6.69 (2.93)
Mechanical Turk 1	6.27 (2.28)	7.74 (4.34)	5.77 (1.17)	5.57 (1.40)	7.14 (3.79)
Mechanical Turk 2	5.76 (1.36)	6.62 (1.85)	5.92 (1.38)	5.78 (1.17)	6.38 (1.75)
Total	5.78 (1.61)	6.94 (2.88)	5.67 (1.18)	5.84 (1.71)	6.57 (2.60)

Table 3

Percent and Average Percent of Frequency for Cue-Feature Part of Speech Combinations

Cue Type	Feature Type	% Raw	% Root	*M* (SD) Freq. Raw	*M* (SD) Freq. Root
Adjective	Adjective	38.09	29.74	17.84 (16.47)	30.02 (18.83)
	Noun	40.02	46.74	13.14 (14.96)	29.71 (19.94)
	Verb	17.69	20.72	8.51 (9.78)	26.88 (17.27)
	Other	4.20	2.80	15.17 (15.64)	28.04 (15.54)
Noun	Adjective	16.56	12.07	15.55 (15.17)	31.20 (18.17)
	Noun	60.85	62.67	17.21 (17.01)	33.26 (20.05)
	Verb	20.80	23.68	8.88 (9.73)	31.01 (17.87)
	Other	1.79	1.58	17.06 (15.29)	28.87 (17.14)
Verb	Adjective	15.16	12.27	13.95 (13.98)	30.03 (18.28)
	Noun	42.92	44.35	14.59 (14.92)	29.59 (18.90)
	Verb	36.92	39.72	12.75 (14.85)	30.43 (19.54)
	Other	5.00	3.66	19.16 (15.95)	25.59 (19.54)
Other	Adjective	20.80	20.32	16.61 (17.37)	31.66 (19.51)
	Noun	42.74	39.03	16.77 (19.41)	37.28 (25.94)
	Verb	19.66	23.93	7.18 (7.57)	26.14 (19.38)
	Other	16.81	16.71	22.72 (16.69)	30.70 (18.48)
Total	Adjective	19.74	14.93	16.12 (15.57)	30.75 (18.37)
	Noun	55.41	57.81	16.55 (16.74)	32.58 (20.09)
	Verb	22.02	24.95	9.50 (10.91)	30.29 (18.24)
	Other	2.82	2.31	17.76 (15.83)	28.45 (16.83)

Note. Raw words indicate original feature listed, while root words indicated translated feature. These data are only from the current project.

Table 4

Example of Affix Coding and Percent of Affixes Found

Affix Type	Example	Percent
Actions/Processes	ion, ment, ble, ate, ize	8.21
Characteristic	y, ous, nt, ful, ive, wise	22.72
Location	under, sub, mid, inter	0.44
Magnitude	er, est, over, super, extra	1.31
Not	less, dis, un, non, in , im, ab	2.76
Number	s, uni, bi, tri, semi	28.31
Opposites/Wrong	mis, anti, de	0.13
Past Tense	ed	8.03
Person/Object	er, or, men, person, ess, ist	7.23
Present Participle	ing	14.03
Slang	bros, bike, bbq, diff, h2o	0.12
Third Person	s	6.16
Time	fore, pre, post, re	0.54

Table 5

Percent and Mean Overlap to the Free Association Norms

	% Overlap	*M* FSG	*SD* FSG	Min	Max	*r*
Adjective	51.86	.12	.15	.01	.94	.36
Noun	36.48	.11	.14	.01	.91	.40
Verb	32.15	.11	.13	.01	.94	.44
Other	44.44	.13	.18	.01	.88	.09
Total	37.47	.11	.14	.01	.94	.39
All Buchanan cues	52.12	.11	.14	.01	.94	.41
McRae et al. cues	23.50	.10	.14	.01	.91	.28
Vinson & Vigliocco cues	15.19	.09	.13	.01	.88	.38
Overlapping Cues	27.26	.09	.14	.01	.88	.30

Note. Overlap was defined as the percent of cue-feature combinations from our feature list included in the Nelson et al. (2004) norms. FSG: Forward strength indicating the number of times a target was elicited after seeing a cue word. Correlation represents the relationship between frequency percent and forward strength.

Table 6

Correlations and 95% CI between Semantic and Associative Variables

	Root	Raw	Affix	PCOS	MV COS	JCN	LSA	FSG	BSG
Root	1	208515	208515	83762	101446	5617	5590	6753	6685
Raw	.93 [.93,.93]	1	208515	83762	101446	5617	5590	6753	6685
Affix	.50 [.50,.50]	.53 [.53,.54]	1	83762	101446	5617	5590	6753	6685
PCOS	.94 [.94,.94]	.91 [.91,.91]	.49 [.48,.49]	1	52342	2762	2759	3280	3243
MV COS	.84 [.84,.84]	.89 [.89,.89]	.46 [.45,.46]	.83 [.82,.83]	1	1179	1179	1248	1232
JCN	-.18 [-.20,-.15]	-.22 [-.25,-.20]	-.17 [-.20,-.15]	-.22 [-.26,-.19]	-.39 [-.44,-.34]	1	5590	5617	5617
LSA	.18 [.16,.21]	.15 [.12,.18]	.10 [.07,.13]	.21 [.18,.25]	.14 [.08,.19]	-.06 [-.08,-.03]	1	5590	5590
FSG	.06 [.04,.08]	.04 [.01,.06]	.08 [.05,.10]	.10 [.06,.13]	.10 [.04,.15]	-.15 [-.18,-.13]	.24 [.22,.27]	1	6685
BSG	.14 [.12,.16]	.15 [.13,.17]	.17 [.14,.19]	.18 [.15,.22]	.26 [.20,.31]	-.18 [-.21,-.16]	.26 [.23,.28]	.31 [.29,.33]	1

Note. Root, raw, and affix cosine values are from the current reprocessed dataset. PCOS indicates the cosine values in the original Buchanan et al. (2013) dataset. MV COS: Cosine values from only the McRae et al. (2005) and Vinson and Vigliocco (2008) data, JCN: Jiang-Conrath semantic distance, LSA: Latent Semantic Analysis score, FSG: Forward Strength, BSG: Backward Strength. Sample sizes for each correlation are presented in the top half of the table.

Table 7

Lexical Decision Response Latencies' Correlation and 95% CI with Semantic and Associative Variables

Variable	First 200	First 1200	Other 200	Other 1200
Root Cosine	.06 [.01,.12]	-.05 [-.10,.01]	.09 [.03,.14]	.09 [.03,.14]
Raw Cosine	.07 [.02,.12]	.05 [-.01,.10]	.09 [.04,.15]	.07 [.01,.12]
Affix Cosine	-.01 [-.06,.05]	.00 [-.05,.06]	.06 [.00,.11]	.04 [-.01,.10]
Target Root FSS	-.02 [-.07,.04]	-.31 [-.36,-.26]	-.03 [-.09,.02]	-.03 [-.08,.03]
Target Raw FSS	-.09 [-.15,-.04]	-.27 [-.32,-.22]	-.03 [-.08,.03]	-.02 [-.08,.03]
Target CSS	-.07 [-.12,-.02]	-.11 [-.16,-.06]	-.05 [-.10,.01]	.02 [-.04,.07]
Cue Root FSS	-.02 [-.07,.04]	-.32 [-.37,-.27]	.03 [-.02,.09]	.03 [-.02,.09]
Cue Raw FSS	.01 [-.04,.07]	-.34 [-.38,-.29]	.01 [-.05,.06]	.01 [-.04,.07]
Cue CSS	.16 [.11,.21]	-.23 [-.28,-.18]	.06 [.01,.12]	.01 [-.05,.06]
Forward Strength	-.12 [-.17,-.06]	-.12 [-.18,-.07]	.07 [.01,.12]	.04 [-.01,.10]
Backward Strength	.15 [.10,.20]	.10 [.04,.15]	.08 [.03,.14]	.04 [-.02,.10]
LSA	.05 [-.00,.11]	-.20 [-.26,-.15]	.13 [.08,.19]	.09 [.03,.14]
Jiang-Conrath	-.05 [-.11,.00]	.11 [.06,.17]	-.05 [-.11,.00]	.01 [-.04,.07]

Note. First indicates first associate, other indicates other associate cue-target relation. 200 and 1200 ms represent the SOA, which is the time from the presentation of the cue to the target. CSS: Cue set size, FSS: Feature set size, LSA: Latent Semantic Analysis distance. Sample size is 1290 cue-target pairs for first associates and 1254 pairs for other associates.

Table 8

Naming Response Latencies' Correlation and 95% CI with Semantic and Associative Variables

Variable	FA 200	FA 1200	OA 200	OA 1200
Root Cosine	-.02 [-.08,.03]	.10 [.05,.15]	-.00 [-.06,.05]	.06 [.00,.11]
Raw Cosine	-.02 [-.07,.04]	.11 [.06,.17]	-.01 [-.06,.05]	.05 [-.01,.10]
Affix Cosine	-.01 [-.07,.04]	.06 [.01,.11]	.03 [-.03,.08]	.01 [-.05,.06]
Target Root FSS	-.03 [-.09,.02]	-.03 [-.09,.02]	-.01 [-.07,.04]	.03 [-.03,.08]
Target Raw FSS	-.04 [-.09,.02]	-.02 [-.07,.04]	-.02 [-.08,.03]	.03 [-.02,.09]
Target CSS	-.06 [-.11,-.00]	-.04 [-.09,.02]	-.02 [-.08,.03]	.01 [-.04,.07]
Cue Root FSS	-.03 [-.09,.02]	-.00 [-.06,.05]	.02 [-.03,.08]	-.02 [-.07,.04]
Cue Raw FSS	-.01 [-.07,.04]	-.01 [-.07,.04]	.02 [-.04,.07]	-.02 [-.07,.04]
Cue CSS	-.01 [-.06,.05]	-.01 [-.07,.04]	-.01 [-.07,.04]	-.01 [-.06,.05]
Forward Strength	-.02 [-.08,.03]	.02 [-.03,.08]	.04 [-.01,.10]	.04 [-.01,.10]
Backward Strength	.10 [.05,.15]	.08 [.02,.13]	.11 [.06,.17]	.04 [-.02,.09]
LSA	.06 [.01,.12]	.03 [-.02,.09]	.06 [.00,.11]	.03 [-.03,.08]
Jiang-Conrath	-.05 [-.11,.00]	.00 [-.05,.06]	-.09 [-.14,-.03]	-.01 [-.06,.05]

Note. First indicates first associate, other indicates other associate cue-target relation. 200 and 1200 ms represent the SOA, which is the time from the presentation of the cue to the target. CSS: Cue set size, FSS: Feature set size, LSA: Latent Semantic Analysis distance. Sample size is 1287 cue-target pairs for first associates and 1249 pairs for other associates.