

eDom: Norming software and relative meaning frequencies for 544 English homonyms

Blair C. Armstrong · Natasha Tokowicz · David C. Plaut

Published online: 5 April 2012
© Psychonomic Society, Inc. 2012

Abstract Words that are homonyms—that is, for which a single written and spoken form is associated with multiple, unrelated interpretations, such as COMPOUND, which can denote an <enclosure> or a <composite> meaning—are an invaluable class of items for studying word and discourse comprehension. When using homonyms as stimuli, it is critical to control for the relative frequencies of each interpretation, because this variable can drastically alter the empirical effects of homonymy. Currently, the standard method for estimating these frequencies is based on the classification of free associates generated for a homonym, but this approach is both assumption-laden and resource-demanding. Here, we outline an alternative norming methodology based on explicit ratings of the relative meaning frequencies of dictionary definitions. To evaluate this method, we collected and analyzed data in a norming study involving 544 English homonyms, using the eDom norming software that we developed for this purpose. Dictionary definitions were generally sufficient to exhaustively cover word meanings, and the methods converged on stable norms with fewer data and less effort on the part of the experimenter. The predictive validity of the norms was demonstrated in analyses of lexical decision data from the English Lexicon Project (Balota et al., *Behavior Research Methods*, 39, 445–459, 2007), and from Armstrong and Plaut (Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, 2223–2228, 2011). On the basis of these

results, our norming method obviates relying on the unsubstantiated assumptions involved in estimating relative meaning frequencies on the basis of classification of free associates. Additional details of the norming procedure, the meaning frequency norms, and the source code, standalone binaries, and user manual for the software are available at <http://edom.cnbc.cmu.edu>.

Keywords Semantic ambiguity · Homonyms · Relative meaning frequency · Norming methods · Rating dictionary definitions · Free associate classification · Homonymy disadvantage

The vast majority of words in English and many other languages are semantically ambiguous—that is, they are associated with multiple distinct interpretations. For example, the word COMPOUND can denote either an <enclosure> or a <composite> meaning, depending on the context (hereafter denoted as <enclosure>/<composite> COMPOUND; Britton, 1978; Klein & Murphy, 2001). As a result, developing an account of semantic ambiguity resolution is a critical component of any theory of word and discourse comprehension, and this has been a key objective in a considerable body of research over the past several decades (e.g., Armstrong & Plaut, 2008, 2011; Frazier & Rayner, 1990; Hino, Pexman, & Lupker, 2006; Hogaboam & Perfetti, 1975; Joordens & Besner, 1994; Klepousniotou, Titone, & Romero, 2008; Mirman, Strauss, Dixon, & Magnuson, 2010; Neill, Hilliard, & Cooper, 1988; Rodd, Gaskell, & Marslen-Wilson, 2002; Rubenstein, Garfield, & Millikan, 1970; Schvaneveldt, Meyer, & Becker, 1976; Simpson, 1984, 1994; Tokowicz & Kroll, 2007). A central challenge in pursuing such work is identifying an appropriate set of ambiguous words to serve as stimuli. One important factor to consider in this process is the relative interpretation frequency, or dominance, of each individual interpretation (Armstrong & Plaut, 2011; Frazier & Rayner, 1990;

B. C. Armstrong (✉) · D. C. Plaut
Department of Psychology and Center for the Neural Basis of Cognition,
Carnegie Mellon University,
5000 Forbes Avenue,
Pittsburgh, PA 15213, USA
e-mail: blairarm@andrew.cmu.edu

N. Tokowicz
Department of Psychology, Learning Research and Development
Center, and Center for the Neural Basis of Cognition,
University of Pittsburgh,
Pittsburgh, PA, USA

Klepousniotou & Baum, 2007; Swinney, 1979; Twilley, Dixon, Taylor, & Clark, 1994).

Dominance plays a key role in the processing of an ambiguous word because it influences the predictability of encountering each interpretation of the word. Prior to the presentation of a context that biases the selection of a particular interpretation, words with relatively balanced meaning frequencies and that lack a clearly dominant interpretation are least predictive of which interpretation should be activated (e.g., the <enclosure> and the <composite> meanings of COMPOUND occur approximately equally often). However, as the meaning frequencies become unbalanced and one meaning becomes clearly dominant, it becomes possible, in principle, to predict which meaning to activate with a high degree of accuracy (e.g., the <lease> meaning of RENT occurs more frequently than the <opening> meaning). As a result, differences in dominance may critically modulate the processing of an ambiguous word by altering the representations that are activated and maintained in neutral contexts (Frazier & Rayner, 1990; Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982; Swinney, 1979; Williams, 1992). When interpretation frequencies are balanced, both interpretations may be partially activated to the same extent. In contrast, when interpretation frequencies are unbalanced and one interpretation is clearly dominant, the dominant interpretation may be strongly activated and the activation of the subordinate interpretations may be substantially reduced (Seidenberg et al., 1982; Swinney, 1979). At the extremes, closely balanced ambiguous words may thus serve as the ideal items for use in many experiments investigating the effects of ambiguity, whereas strongly unbalanced ambiguous words may be virtually indistinguishable from the unambiguous words with which they are typically contrasted (Armstrong & Plaut, 2008; Hino et al., 2006; Klepousniotou & Baum, 2007; Rodd et al., 2002).

The difference between balanced and unbalanced ambiguous words is most pronounced in the case of homonyms—words for which a single written and spoken form is associated with multiple *unrelated* interpretations, and for which there is general agreement that the semantic overlap between the interpretations is minimal (e.g., <dog>/<tree> BARK; Armstrong & Plaut, 2008; Frazier & Rayner, 1990; Hino, Kusunose, & Lupker, 2010; Hino et al., 2006; Klein & Murphy, 2001, 2002; Klepousniotou et al., 2008; Rodd et al., 2002; Rodd, Gaskell, & Marslen-Wilson, 2004; Rubenstein et al., 1970). This contrasts with *polysemes*, for which a single written and spoken form is associated with multiple *related* interpretations, which may reduce the degree to which each individual meaning may be differentially activated (Armstrong & Plaut, 2008, 2011; Beretta, Fiorentino, & Poeppel, 2005; Frazier & Rayner, 1990; Klepousniotou et al., 2008; Pytkänen, Llinás, & Murphy, 2006; Rodd et al., 2002; but see Hino et al., 2010;

Hino et al., 2006; Klein & Murphy, 2001, 2002, for dissenting views). Consequently, assessing the dominance of homonyms is particularly important, and failures to control for this factor have been proposed as an explanation for the weak and inconsistent effects of homonymy in many studies (Armstrong & Plaut, 2011).

The goals of the present work were threefold: (1) to develop an efficient and reliable technique for estimating the meaning frequencies of homonyms primarily on the basis of ratings of dictionary definitions, which avoids several problems with classic methods for assessing meaning frequency; (2) to collect normative data for a pool of words, suitable for use in future investigations, that could also be used to examine the characteristics of the norms and the reliability of the method; and (3) to demonstrate the predictive validity and utility of the norms by analyzing the results of lexical decision experiments reported previously by Armstrong and Plaut (2011) and as part of the English Lexicon Project (Balota et al., 2007). In realizing these goals, we have developed open-source software to automate the norming process and have collected norms for 544 homonyms that we are making available for use and extension by other researchers. This article presents a brief overview of this work. Additional details, as well as the normative data and software, are available in the online user manual, located at <http://edom.cnbc.cmu.edu>.

Issues with existing norming methods

One popular method for estimating dominance is via the classification of the free associates generated for a given homonym on the basis of the meaning of the word to which they are related (Geis & Winograd, 1974; Gilhooly & Logie, 1980a, 1980b; Gorfein, Viviani, & Leddo, 1982; Kausler & Kollasch, 1970; Mirman et al., 2010; Nelson, McEvoy, Walling, & Wheeler, 1980; Twilley et al., 1994), which is related to similar methods of classifying generated definitions (Warren, Bresnick, & Green, 1977), generated sentences (Wollen, Cox, Coahran, Shea, & Kirby, 1980), or sentence completions (Yates, 1978). These methods involve two steps: (1) Participants are provided with an ambiguous word (e.g., BANK) and generate an associate (or other similar response; e.g., MONEY), and (2) a separate group of raters classify these responses on the basis of their intuitions regarding the meanings to which these associates are related (e.g., the <financial> vs. <edge of a river> meanings of BANK). Across a number of studies, this method has been shown to generate a fairly consistent measure of dominance (for a review, see Twilley et al., 1994).

There are, however, several issues with these methods. At a theoretical level, researchers must make the assumption that responses in the free-association task are generated in direct proportion to the relative frequencies of each of a word's meanings. Although there may be some surface validity to

this claim, correlations between norms generated via the free-association technique, although they are high among other studies using the same method, generally decrease when other, similar techniques, such as the classification of generated sentences or definitions, are used (Twilley et al., 1994). This suggests that a nontrivial task-specific component is involved in the ratings generated by classifying free associates.

Bridging between the theoretical and methodological levels, raters often encounter difficulty in agreeing on which meaning each free associate should be linked with, if any; the average overlap in rater classifications in Twilley et al.'s (1994) study was only between 65 % and 75 %. Participants often produce associates that are not strongly semantically related to either interpretation of a homonym, which makes it difficult to establish consistent classifications across raters. This is true even under the assumption that the participants and raters have identical semantic representations. For instance, if associations are weak, random noise in each rater's classification process may prevent consistent classification across raters. However, this assumption may not be valid, because there may be differences in both the quantity and the type of discourse to which the raters and the participants have been exposed, which may in turn cause each group to develop somewhat different semantic representations. At the very least, the low agreement across raters puts in question the efficacy of using data from free-association tasks to generate relative meaning frequency ratings, given the low information content of each associate/rating. At worst, it suggests that other, nonsemantic types of associations may influence response generation to a substantial degree. An examination of the responses from free-association tasks supports the latter conclusion: Responses often consist of synonyms (e.g., COP \Rightarrow OFFICER), antonyms (e.g., HOT \Rightarrow COLD), category coordinates (e.g., ROBIN \Rightarrow SPARROW), completions that form compound words (e.g., WRIST \Rightarrow WATCH), and other associates that are not purely semantic in nature. These results invalidate the simple assumption that responses are semantic associates of the target and are generated as a function of the target's relative meaning frequencies. Clearly, a more refined theory of how responses are generated in free-association tasks is needed to improve the relative meaning frequency estimates that can be extracted from free-association norms. It is, therefore, worth developing a norming methodology that avoids these issues.

Another issue with the classification of free associates is how raters agree on the initial set of word meanings into which each response should be classified. To increase the consistency of this process, researchers often classify the associates into the meanings of the word listed in a dictionary (e.g., Mirman et al., 2010; Twilley et al., 1994). However, no evidence has been provided to show that dictionary definitions are sufficiently similar to the mental representations of word meanings

in the target populations to be suitable for this task. For example, dictionary definitions may fail to include representations of vernacular meanings and may contain meanings based on the etymology of the word that are no longer in common usage, which may hinder the extraction of accurate meaning frequency estimates.

Finally, these techniques are resource-intensive in several respects. The average number of participants that must generate associates is typically quite large in these studies, with many studies collecting ratings from well over 100 participants for only about 100 homonyms (Twilley et al., 1994). Additionally, each participant's responses must be classified individually by one, or preferably several, raters. With each study typically involving over 100 words, each rater must therefore classify over 10,000 observations. The end result of these resource demands is that large-scale norming studies are rare. Given continued concerns that a word's relative meaning frequencies can differ substantially across different participant populations and can change over time, given the fluid nature of language (Swinney, 1979; Twilley et al., 1994), an alternative method that makes such norming studies more tractable is clearly desirable.

An alternative method: Rating dictionary definitions

The present study investigated the reliability, validity, and efficiency of norms based on explicit relative meaning frequency ratings of dictionary definitions (supplemented by participant-generated definitions) in a set of homonyms. From a theoretical standpoint, this method is a more direct assay of meaning frequency because it avoids the need to make assumptions about how the associates were generated. Furthermore, it avoids the methodological drawbacks of investing considerable resources in having raters classify participants' responses (often according to the definitions listed in a dictionary) and of the inconsistent classifications that often arise during this process across raters. Finally, the directness of this method (one participant \Rightarrow one dominance rating vs. many participants \Rightarrow classification of many responses \Rightarrow one dominance rating) should also lead to more rapid convergence on stable norms.

Method

Participants

A total of 64 (24 male, 40 female) native English speakers, 18 years of age and above, who were enrolled in psychology courses at the University of Pittsburgh participated in the experiment in exchange for course credit. There was no explicit screening to exclude participants with language disorders, although none were spontaneously reported by the participants.

Stimuli

Homonyms and a small number of homographs (i.e., words with a single orthographic form but two phonological forms associated with two different meanings; e.g., <turn>/<storm> WIND) that would be ideally suited for standard semantic ambiguity experiments were selected for norming on the basis of the standard parameters of several variables (see, e.g., Armstrong & Plaut, 2011; Rodd et al., 2002). Specifically, these stimuli consisted of all words between 3 and 10 letters in length, with a \log_{10} (SUBTL word frequency) between 1 and 100 (Brysbaert & New, 2009),¹ with sense counts in wordNet (Fellbaum, 1998), with phoneme and syllable counts in N-Watch (Davis, 2005), and with two or more unrelated meanings in the Wordsmyth online dictionary, which has been employed in past semantic ambiguity studies (e.g., Armstrong & Plaut, 2011; Azuma & Van Orden, 1997; Rodd et al., 2002) and is available online at wordsmyth.net (Parks, Ray, & Bland, 1998). According to the classification scheme used in the online version of Wordsmyth, separate webpage entries denote unrelated meanings of a word, whereas separate definitions on a single page denote distinct but related senses of a word. Furthermore, the order in which each meaning appears reflects the rank-ordered frequency of that meaning according to the Wordsmyth lexicographers. This coarse meaning/sense classification correlates with estimates of homonymy and polysemy (Azuma & Van Orden, 1997; Rodd et al., 2002). A total of 585 words satisfied these constraints—576 homonyms and 9 homographs, which were collapsed in with the homonyms in all of the remaining analyses—of which 483 had two meanings, 84 had three meanings, 15 had four meanings, and 3 had five meanings. A parsing script was used to extract the definitions for each of these words from Wordsmyth.² Given the large number of words that were normed, as well as the screening criteria employed for selecting these words, the results of the present norming study are broadly representative of the population of words that could appear in a typical experiment that employs homonymous stimuli.

Procedure

Due to the size of the word set, it was not feasible to have each participant rate each word in a single session. Instead, each participant rated a random sample of approximately one quarter of the full set, or 146 words, resulting in approximately 16 ratings per word in total. To ensure that the words were seen by equal numbers of raters, these samples

were generated by randomly sampling from the population without replacement until the population was exhausted, at which point the population was reset and the process was repeated. The randomization script that accomplished this is included with the eDom software.

Before beginning the experiment, participants were instructed that they would be estimating, as a percentage, how often a particular meaning of a presented word was implied when they encountered that word, and they were given examples of a balanced and an unbalanced homonym. They were told that the dictionary definitions of the words were listed to remind them of the meanings associated with the word, and they were instructed to read over the definitions to determine which meaning should be estimated. They were told that their estimates, however, should be based on their own personal experience and that the number, order, and length of the definitions should not directly impact their judgments. Additionally, they were instructed to list up to two additional meanings of a given word if they knew of meanings that were not included in the presented definitions. If additional meanings were listed, the participant's relative meaning frequency ratings were to include ratings for those extra meanings. If participants did not know a word at all, they were to respond "don't know." They were also instructed to try to be accurate in their ratings without spending too much time thinking about how to rate a particular word. The participants were prompted to ask any questions that they had for the experimenter prior to the beginning the experiment. The full instructions used in the experiment are available in the online manual.

A custom application called eDom was created to present the words and their definitions to participants. The full details of this software, as well as the source code and standalone binaries for several operating systems, are available to researchers via the online user manual. In brief, participants were presented with a 3×2 array in which each cell (left to right and top to bottom) contained all of the definitions of the related senses associated with a distinct meaning (see Fig. 1). The order in which the definitions associated with each meaning were presented was randomized so as to avoid presenting dominant and subordinate meanings on the basis of their order of entry in Wordsmyth. Additionally, two of the cells were editable and shaded yellow, and these were available for participants to list other definitions of the words that they knew (only one such cell was available for the three stimuli with five meanings). Below each meaning was a field into which participants could enter their estimate of the meaning frequency, as a percentage of the time that that meaning was implied when the word was encountered (default value = 0). New definitions listed by participants were required to have nonzero percentages, and the sum of all of the percentages had to be 100. Once participants were done rating a word, they

¹ This upper bound also avoids the need to control for an additional quadratic frequency component due to a ceiling effect (Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Brysbaert & New, 2009).

² Five words were excluded because the automated parser did not correctly extract their definitions from Wordsmyth or there was a duplicate entry in the online dictionary.

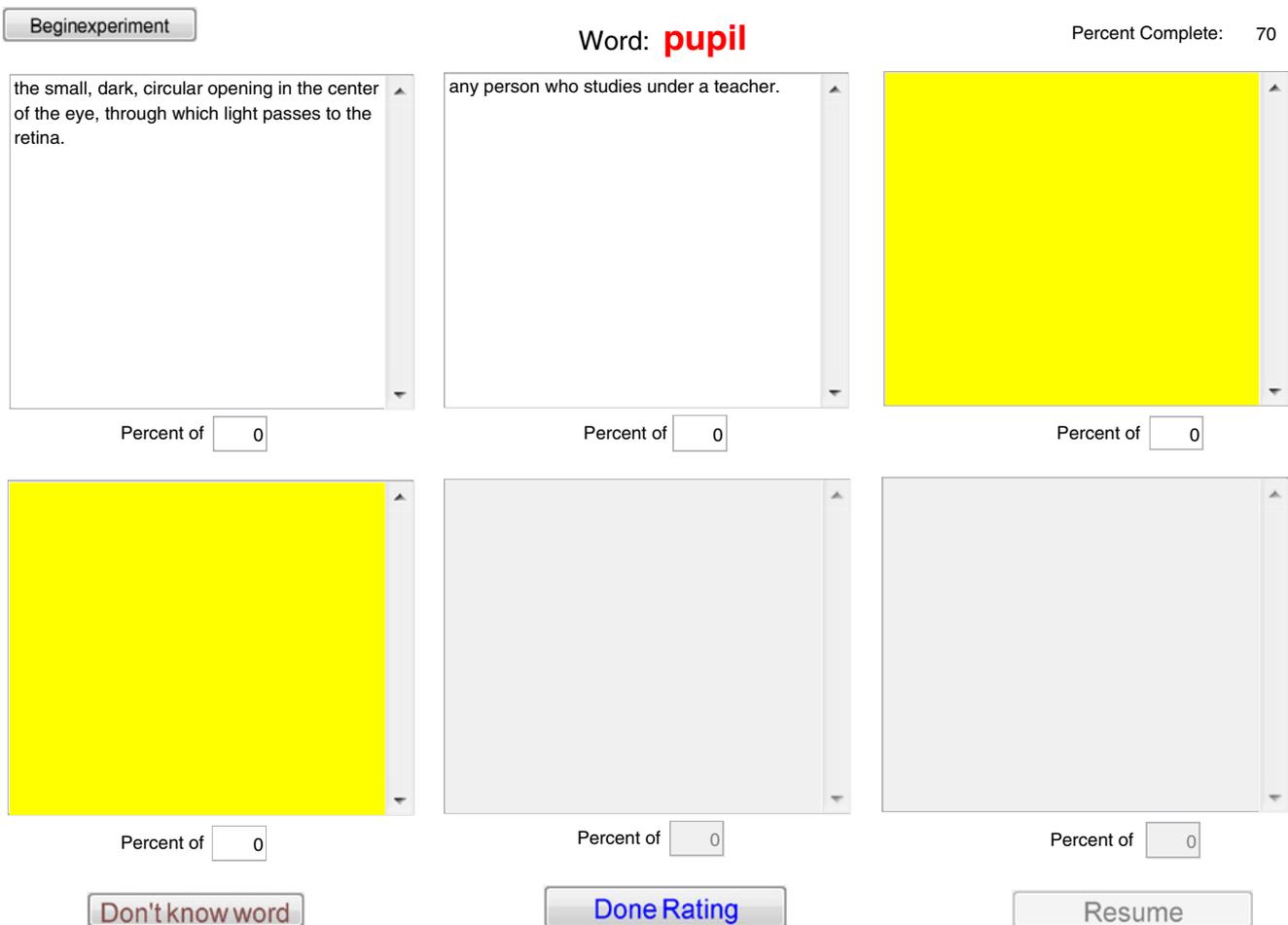


Fig. 1 Screen capture of the eDom norming software during the norming of the word PUPIL

pressed a “done rating” button and a new word was presented. Alternatively, if participants did not know any of the meanings of a word, they could press the “don’t know” button. Self-paced breaks were available after every 24 words. The entire experiment required approximately 50 min to complete.

Results

Removing data for unknown words

On average, participants responded that they did not know a word on 3 % of trials. In total, there were 32 words for which at least approximately 20 % of respondents (three participants) responded “don’t know”; these responses corresponded to 2.3 % of the total responses. These words were excluded from further analysis, along with all of the other trials on which a “don’t know” response was made. This left 553 words in the trimmed set, of which 544 were homonyms.

Participant supplementation of dictionary definitions

The participants listed additional meanings for the presented words on 2.6 % of the trials. These responses were distributed approximately equally across each block of the experiment. Twenty words had additional definitions given by at least 20 % of the respondents (three participants); the additional definitions for these words accounted for approximately 34 % of all the participant-generated definitions. One of these words was part of the set that was removed due to exceeding the “unknown word” criterion, as detailed above. The first author manually reviewed each of the new definitions that participants listed for these 20 words and noted the existence of a consistent new meaning each time that at least 20 % of the respondents’ definitions related to the same new meaning. In total, consistent new meanings were identified for 11 words, and the majority of these definitions were related to vernacular meanings of the words (e.g., <muscular> BUFF). A list of these extra meanings is available in the online manual. The average relative meaning frequency of the participant-generated definitions was 42 %, and only 24 % of these ratings had relative meaning frequencies greater than 50 %. Taken

together, the relative paucity of participant-generated definitions, combined with the moderate convergence of the participants' definitions on specific meanings, suggests that dictionary definitions are relatively successful at exhaustively capturing the meanings of most words. Additionally, in cases in which the dictionary definitions did not represent a common meaning of a word, the participants supplemented the dictionary definitions with their own definitions.

Comparison of the meanings listed in the dictionary with those that were familiar to the participants

On average, the sum of the relative meaning frequencies across all of the dictionary definitions accounted for 99 % of the sum of the relative meaning frequencies across all of the words' dictionary and participant-supplemented definitions. This indicates that dictionary definitions come very close to constituting an exhaustive list of the meanings of homonyms. To determine whether some of these meanings were listed in the dictionary purely for etymological reasons and were not generally encountered by the participants, the average relative-frequency ratings for the dictionary definitions were rank ordered, and the summed ratings across the first through the second, third, fourth, and fifth dictionary meanings were calculated, where applicable. Summing across the two most frequent meanings accounted for the clear majority of encounters with a word (80 % of encounters for words with four or five meanings, 90 % of encounters for words with three meanings, and 99 % of encounters for words with two meanings). These results imply that a number of the meanings associated with the words are either unknown to participants or are known but virtually never encountered. They also indicate that the bulk of encounters are captured by only the first and second meanings of each word, which allows for considerable simplification of the metric used to assess the dominance of a homonym (Twilley et al., 1994). On the basis of this observation, in the following sections of the article we will examine the characteristics of the normed words on a very simple measure of dominance—the highest meaning frequency associated with each word, denoted β for *biggest*. Using the approximation that the sum of the relative meaning frequency ratings for the first two meanings captures all of the relative-frequency data for all of the word's meanings, this measure is effectively a difference score between the first and second most frequent meanings.³

³ This evidence notwithstanding, the correlations between β and other, more complex measures of dominance—such as the B measure introduced by Twilley et al. (1994), which is based on information-content theory, and an alternative measure, D , which we developed on the basis of a standardized difference between the highest and second highest meaning frequencies [(highest – second highest)/highest]—were nevertheless very high (all r s $\leq .94$). Similar results would therefore be expected if using these other measures.

In addition to calculating the overall degrees to which dictionary definitions span participants' representations of a word's meanings, we also examined whether the rank ordering of each definition of a word in Wordsmyth agreed with the rank ordering of participants' meaning frequency estimates. To do so, we assigned a score of 1 each time the rank ordering of the two most frequent meanings in each participant's ratings corresponded to the same rank ordering of the dictionary definitions, and 0 otherwise. We then collapsed these ratings across participants and tested whether the resulting scores tended to agree with the participant ratings, using a single-sample t test with a null hypothesis corresponding to chance agreement (50 %). The dictionary and participant ratings agreed 81 % of the time on average, which was significantly higher than chance, $t(552) = 27$, $SE = 0.01$, $p < .0001$.

Characteristics of the norms and implications for studies that employ homonymous stimuli

Reliability To evaluate the reliability of the ratings that we collected, we first computed the mean β for each word in the norms and then correlated each participant's ratings of a subset of these words with the mean ratings for the averaged item data. We found reasonable consistency between individual participants and the mean ratings ($r = .70$, $SE = .01$, range = .41 – .85). This consistency is comparable to that reported in a similar task in which participants rated the age of acquisition of each meaning of an ambiguous word (Khanna & Cortese, 2011). This provides initial support for the reliability of these norms and of their relative invariance to the particular set of words that form the context in which participants generate ratings.

Next, we examined whether averaging across the data from 16 participants per word, when each participant rated a different subset of the entire set of words, was sufficient to obtain stable estimates of β . To do so, we compared the average β for each homonym obtained in the present norming study to the β data collected in another study using the same norming method (Armstrong & Plaut, 2011). In that study, 50 participants normed each of 200 words with multiple meanings used in the experiment reported therein, of which 195 overlapped with those in the set of words from the present study. Thus, the norms collected as part of the Armstrong and Plaut (2011) study were derived from more than three times as many observations (50 vs. 16) as those in the present study. The correlation between the mean β values across these two sets was nevertheless very high ($r = .95$). Additionally, when the β ratings from the 16-participant data set were used to predict the β ratings from the 50-participant data set in a linear regression, the coefficient for the intercept was near zero ($b = 5.8$, $SE = 1.8$), and the coefficient for the 16-participant β was near 1 ($b = 0.95$, $SE = 0.02$). This confirms that the raw values from each norming experiment, and not simply the

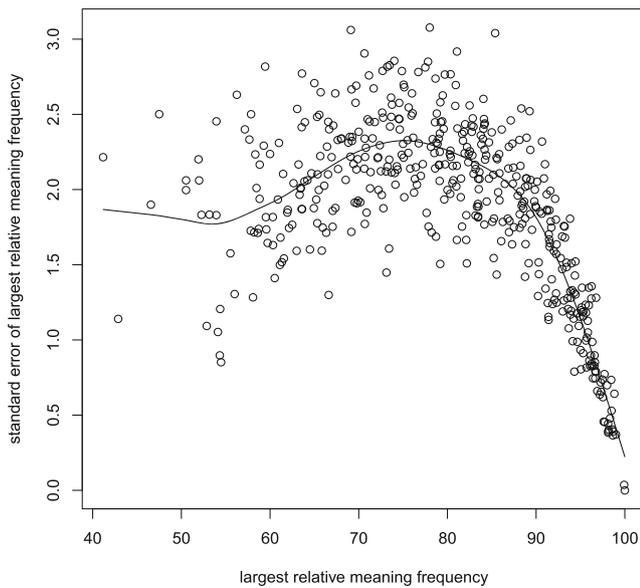


Fig. 2 Standard errors of the largest meaning frequency estimates for each normed word as a function of each word's largest meaning frequency. The solid line depicts the mean of the standard errors for all words with the same largest meaning frequency, fit with a cubic smoothing spline

linear relationship between these values, were highly similar. Taken together, these results indicate that stable ratings were obtained from the 16 participants for each stimulus in the present study, and that the present norming procedure converges on reliable norms more rapidly and with fewer observations than do other norming methodologies. Consequently, there is little to be gained by collecting data from more participants or from requiring participants to rate the entire set of words. This contrasts with the original arguments Twilley et al. (1994) used to motivate their large-scale free-association norming study.

Additionally, we examined the reliability of the β measure as a function of the value of β by plotting the standard error of β as a function of the β for each word⁴ (see Fig. 2). Contrary to past research suggesting that the most balanced words are associated with the least stable ratings (Swinney, 1979; Twilley et al., 1994), we observed moderate variability for β values between 40 and 60, the highest variability between 60 and 90, and rapidly decreasing variability between 90 and 100. The lack of a reduction in reliability for words with β values less than about 90 also provides support for our β measure as an alternative to more complex dominance metrics. For instance, our data do not support using a metric based on information-content theory, such as the

⁴ A second plot, not included in this article, established that the relationship depicted in Fig. 2 was not due to a confound with several other variables. This was accomplished by plotting “residual β ,” for which the correlated contributions from all of the variables listed in the Correlation With Other Variables section had been removed. The resulting plot produced a result qualitatively similar to that presented here.

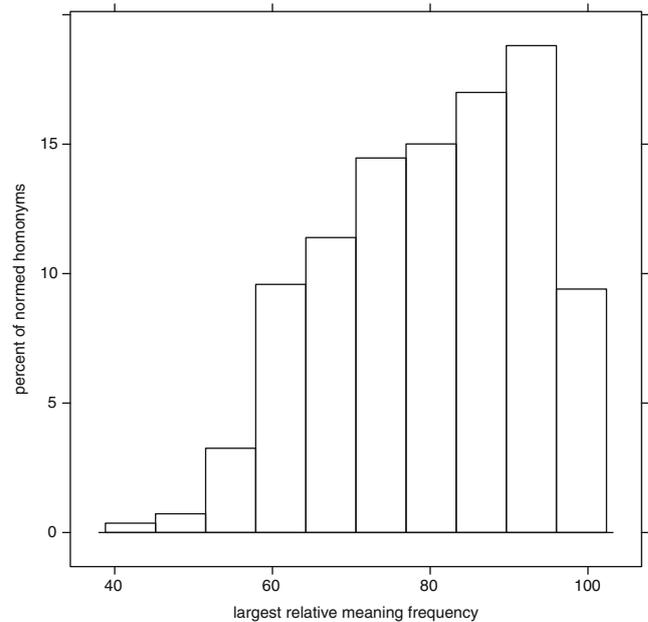


Fig. 3 Histogram of the distribution of the largest relative meaning frequency values for all of the normed words

uncertainty measure U derived by Twilley et al., which loses sensitivity as words become more balanced.

Distribution of largest relative meaning frequency Examining the distribution of β values provides insight into the expected characteristics of experimental words that have not controlled for relative meaning frequency during item selection. A histogram of the β scores for all of the normed homonyms is presented in Fig. 3. This figure shows that these data are moderately left-skewed and that the majority of the homonyms (~65 %) have their largest relative meaning frequencies in excess of 75 %. Given that several studies have defined a homonym as having relatively balanced meaning frequencies if the largest relative meaning frequency was less than 75 %, or an even more conservative value (Armstrong & Plaut, 2011; Klepousniotou & Baum, 2007; Mirman et al., 2010; Swinney, 1979), this indicates that a random sample of words from our relatively exhaustive norming of homonyms in English would fail to satisfy this constraint. This finding emphasizes the need to constrain item selection on the basis of relative meaning frequency a priori, or otherwise to address the effects of this variable—for instance, by including it as a covariate when analyzing the data—as well as the utility of our norms for this end. This distribution may also help explain the small magnitude of the homonymy disadvantage that has been reported in several studies (Armstrong & Plaut, 2008, 2011; Beretta et al., 2005; Rodd et al., 2002). In many of these experiments, no consideration of dominance was used in selecting words, and as a result, over 50 % of the words in those experiments had largest relative meaning frequencies above 75 % (Armstrong & Plaut, 2011). Consequently, the weak or nonexistent homonymy

effects in these experiments may have been due, at least in part, to the fact that the majority of the words were closer to being unambiguous than they were to being optimally ambiguous.

Correlation with other variables Using correlation analyses, we assessed whether β was significantly related to other common variables that are controlled for or experimentally manipulated in studies involving homonymous stimuli. These correlations were significant ($p < .05$) when the correlation coefficient was greater than .10, given that there were 553 observations entered into each statistical test. Marginal effects ($p < .10$) are also indicated in the discussion below.

At the semantic level, we observed significant correlations with the number of unrelated meanings ($r = -.32$) and the number of related senses ($r = -.28$), as well as with the number-of-interpretations⁵ counts in wordNet ($r = -.27$). The correlations with these measures emphasize the need to carefully consider how to design studies that can detect strong main effects of relative meaning frequency, given that the number of meanings and the number of senses associated with a word have been linked with processing disadvantages and advantages, respectively (Armstrong & Plaut, 2008, 2011; Klepousniotou & Baum, 2007; Rodd et al., 2002).

At a grammatical level, correlations with β and the number of distinct interpretations falling into particular grammatical classes, as listed in Wordsmyth and collapsed across meanings and senses, were observed for verbs ($r = -.26$) and nouns ($r = -.16$). No significant correlations were detected for adjectives ($r = -.07$) or adverbs ($r = -.05$). These data reveal that grammatical class is another important property to consider when using homonymous stimuli to study word and discourse comprehension, particularly with respect to nouns and verbs. It also opens the possibility that the different ambiguity effects that have been reported across grammatical classes are the result of a confound with relative meaning frequency (Frazier & Rayner, 1990; Frisson & Pickering, 1999; Pickering & Frisson, 2001), although Mirman et al. (2010) observed different effects for noun–noun versus noun–verb homonyms, even when controlling for a measure of relative meaning frequency.

At a lexical level, there were significant but weak correlations with both raw and log-transformed word frequency ($r_s = -.11$), which agree with several past studies showing that such a relationship, if it exists, is weak (for discussion, see Twilley et al., 1994). The correlation with orthographic Levenshtein distance (Yarkoni, Balota, & Yap, 2008) was also significant ($r = .14$), and the correlation with word length (in number of letters) was marginally significant

($r = .10$). No significant correlations were detected with sublexical measures of number of phonemes ($r = .04$), number of syllables ($r = .08$), or positional bigram frequency ($r = .09$), which supports the assumption that sublexical and semantic representations are independent (Armstrong & Plaut, 2008; Plaut, 1997).

Overall, these significant correlations point to the importance of carefully controlling or otherwise addressing these potential confounds when designing studies that investigate the effects of homonymy, relative meaning frequency, and many of the other factors listed above. The large number of significant correlations also suggests that automated methods of eliminating these confounds (e.g., Armstrong, Watson, & Plaut, 2012) may be particularly valuable in such undertakings.

Generalizability

Comparison to relative-frequency estimates derived from the classification of free associates To evaluate the degree to which our method of assessing dominance taps the same underlying representations as norms based on the classification of free associates, we correlated β with data from a large-scale norming study conducted using the latter methodology (Twilley et al., 1994). The Twilley et al., study reported a measure referred to as U , which consisted of a nonlinear transformation of the relative frequencies for each meaning. However, the difference between expressing a word's relative meaning frequencies either as β or as U is very small for words with two meanings. This was determined by correlating these two measures across 21 equally spaced data points that spanned the range of possible values of β/U for words with two meanings (i.e., β values in the range [.5, 1]). This yielded an r of .94. On the basis of this result and the fact that the bulk of the words in the present study were reported as essentially having only two meanings by our participants, we therefore assumed that the differences observed when comparing β from our norms to U from Twilley et al. are primarily due to task differences. After adjusting for differences in the interpretation of the sign of the slope of each measure— β scores increase as a homonym's meanings become less balanced, whereas the opposite is true for U —the two measures correlated only weakly ($r = .27$). This result indicates that although these two measures tap a common underlying variable, for the most part they measure unique variance. Qualitatively, this is in line with the decrease in correlation between measures of relative meaning frequency observed across free-associate classification, definition writing, and sentence generation (see Twilley et al., 1994). Nevertheless, the correlations between the Twilley et al., data and the data obtained using those alternative methodologies were considerably stronger (smallest $r = .72$). Additional work is

⁵ Typically, these are referred to as “sense counts,” but the term “interpretation” is used here to emphasize that these data are an aggregate measure of the number of both unrelated meanings and related senses.

needed to more fully understand the cause of these differences. The comparison of the predictive validity of norms generated by each of these methods, presented next, provides guidance to this end.

Predictive validity of the highest relative meaning frequency

Analysis of the data reported by Armstrong and Plaut (2011) Several previous studies have investigated how homonymy influences word comprehension using the visual lexical decision task (e.g., Armstrong & Plaut, 2008; Azuma & Van Orden, 1997; Klepousniotou & Baum, 2007; Rodd et al., 2002). These studies suggested that once the task is made difficult—for instance, by matching the orthographic characteristics of the words and nonwords to reduce the informativeness of orthography for selecting a response—homonyms show a processing disadvantage relative to unambiguous controls (see Armstrong & Plaut, 2008, for a mechanistic account of this effect). However, this processing disadvantage has typically been weak and has often failed to reach statistical significance ($p < .05$). Armstrong and Plaut (2011) hypothesized that this was in part attributable to failing to control for relative meaning frequency—an issue that was discovered in post hoc norming of the words used in a previous study. Additionally, they hypothesized that these effects would be stronger if even more difficult versions of the lexical decision task were employed.

To address these issues, Armstrong and Plaut (2011) examined whether task difficulty influences the magnitude of the homonymy effect by testing this effect with the same word stimuli in four conditions formed by crossing nonword difficulty (easy vs. difficult; here, we do not consider the preliminary findings reported in that article, concerning a third nonword condition in which pseudohomophones were presented) and stimulus contrast (i.e., presenting words as white-on-black vs. gray-on-black: high contrast vs. low contrast). Because the high-contrast conditions yielded no significant homonymy effects in that study, here we report a reanalysis based on only the low-contrast conditions. We specifically focus our analyses on stimuli that were labeled as (relatively) “unambiguous controls” and “homonyms” in that study. Because it will be relevant later, we also note that these stimuli were further constrained such that there were few related senses associated with each of the meanings of the homonyms and unambiguous controls. This somewhat relaxed definition of what constitutes a “homonym” was necessary because “pure” homonyms that have only distinct, unrelated interpretations are extremely rare—there are only 38 such words in the norms reported here, and only eight of these words have β scores below 75 %. Critically, however, the unambiguous controls and homonyms were matched to have the same total number of related senses, summed across all of their interpretations. Consequently,

these conditions differ only in terms of whether the interpretations are clustered as a single set of related interpretations or are spread across two distinct interpretations. Differences between these conditions can therefore be attributed to whether the words were homonymous or not. Given that relative-frequency norms were not available for a large set of stimuli when Armstrong and Plaut selected words for use in their experiment, they did not control for this factor when selecting their experimental stimuli. Instead, they made the a priori decision to norm the homonyms they selected using the method reported in this article while collecting data in the lexical decision task, and to control for relative meaning frequency during later analyses.

We compared the homonyms and unambiguous controls from each of the easy and difficult lexical decision experiments in three different analyses. In the first analysis, we entered Number of Meanings (one vs. many) as a factor and analyzed the full data set. In the second analysis, we also entered Number of Meanings as a factor, but restricted the “homonym” condition to include only the data for 14 well-balanced homonyms. These homonyms had a mean β of 59 %, and all of the homonyms had a β less than or equal to 65 %. This operational definition of a “balanced” homonym is similar to that employed by Klepousniotou and Baum (2007). We also found essentially the same results using an upper bound of 75 %, as in Mirman et al. (2010), which allowed for approximately double the number of homonyms to be entered into the analysis. In the third analysis, instead of entering Number of Meanings as a factor, we entered β as a continuous variable and analyzed the full data set (unambiguous controls were assigned a β of 100). These are referred to as the *full-factor*, *restricted-factor*, and *full-regression* analyses. All of the analyses were conducted using mixed-effect regression, and p values for the statistical tests were calculated via Monte Carlo simulations (Baayen, Davidson, & Bates, 2008). In addition to the Number of Meanings factor or β variables, these analyses also included \log_{10} (SUBTL word frequency), word length (number of letters), orthographic Levenshtein distance, number of phonemes, number of syllables, residual familiarity,⁶ trial rank, and the lexicality, accuracy, and latency of the previous trial (Baayen & Milin, 2010) as fixed effects, and participants and words as random effects. All of these fixed and random effects succeeded in predicting significant amounts of variance in at least a subset of the analyses conducted by Armstrong and Plaut (2011), and additional predictors such as imageability and positional bigram frequency

⁶ Residual familiarity was calculated by regressing out the effects of the Number of Meanings and Frequency factors from the raw familiarity scores. As noted in Armstrong and Plaut (2011), these two measures correlated very strongly ($r = .98$), and so, for consistency with the analyses in previous work, residual familiarity was used in all of the present analyses.

were not included, because they did not predict significant amounts of variance in Armstrong and Plaut's analyses. Only correct responses were included in the latency analyses, which were measured in milliseconds. For ease of interpretation, in the remainder of this article the slopes for the different coefficients have been standardized, such that a positive slope in accuracy and a negative slope in latency always indicate a homonymy disadvantage.

In the full-factor analyses, no significant effects of homonymy were observed (for easy nonwords: accuracy, $b = 0.0001$, $SE = 0.0083$, $p = .99$, $n = 9,933$; latency, $b = -6.8$, $SE = 4.1$, $p = .09$, $n = 9,129$; for hard nonwords: accuracy, $b = 0.0115$, $SE = 0.0096$, $p = .23$, $n = 10,439$; latency, $b = 1.2$, $SE = 4.4$, $p = .78$, $n = 9,562$). However, a significant or marginal homonymy disadvantage was observed in all of the latency analyses that included some consideration of the effects of relative meaning frequency, with the exception of the "hard" nonword condition in the restricted-factor analysis (for easy nonwords: restricted factor, $b = -16.7$, $SE = 8.3$, $p = .04$, $n = 5,301$; full regression, $b = -0.35$, $SE = 0.16$, $p = .02$, $n = 9,129$; for hard nonwords: restricted factor, $b = -7.6$, $SE = 9.0$, $p = .37$, $n = 5,597$; full regression, $b = -0.29$, $SE = 0.18$, $p = .09$, $n = 9,562$). Several effects approached significance in the accuracy analyses, as well (for easy nonwords: restricted factor, $b = 0.0326$, $SE = 0.0168$, $p = .05$, $n = 5,818$; full regression, $b = 0.0002$, $SE = 0.0003$, $p = .55$, $n = 9,933$; for hard nonwords: restricted factor, $b = 0.0257$, $SE = 0.0162$, $p = .12$, $n = 6,110$; full regression, $b = 0.0006$, $SE = 0.0004$, $p = .10$, $n = 10,439$). We defer further discussion of these results until after presenting the analyses of a second set of lexical decision data.

Using β and U to predict lexical decision performance in the English Lexicon Project To further assess the predictive validity of β , and to directly compare the predictiveness of β to that of U , we evaluated how well each measure predicted performance in the lexical decision task⁷ conducted as part of the English Lexicon Project (Balota et al., 2007). In this lexical decision task, words were presented at full contrast, and much easier nonwords were used than in Armstrong and Plaut (2011); together, these differences in procedure may alter the effect of homonymy. Indeed, data from the English Lexicon Project (although not from the specific set of words analyzed here) were reported recently to show a homonymy advantage, even when smaller-scale lexical decision experiments have failed to reach significance on these comparisons (Hargreaves, Pexman, Pittman, & Goodyear, 2011). The exact cause of these opposing effects of homonymy is an unresolved

issue in the literature and beyond the scope of the present work (but see Hargreaves et al., 2011; Hino & Lupker, 1996; Hino et al., 2006; Kawamoto, 1993, for several related accounts). Here, we focus on whether different measures of relative meaning frequency can contribute to understanding these effects by increasing the magnitude and reliability of the homonymy advantage reported using this data set.

The English Lexicon Project contained data for 551 of the normed words in our data set and for 211 words that appeared in both our data set and the Twilley et al. (1994) norms. Given the substantial differences in the number of available words, we conducted two sets of analyses. The first included only the words for which the U measure was available, so as to compare both β and U on an equal footing, and the second included the full data set, to maximize the generalizability and statistical power of the analysis. For brevity, the different analyses involving β are referred to via different subscripts on the β coefficients: f for the *full* data set and u for the data set for which the U measure was available.

First, we used β and U to predict accuracy and latency using simple regression. These analyses showed that β was a significant or marginal predictor of both measures of performance (accuracy: β_f , $b = -0.0006$, $SE = 0.0003$, $p = .01$; β_u , $b = -0.0004$, $SE = 0.0002$, $p = .05$; latency (in milliseconds): β_f , $b = 0.4$, $SE = 0.2$, $p = .07$; β_u , $b = 0.6$, $SE = 0.3$, $p = .05$), but that U was not (accuracy: $b = 0.001$, $SE = 0.006$, $p = .77$; latency: $b = -13$, $SE = 8$, $p = .10$). Specifically, β predicted higher accuracies and shorter latencies for more balanced homonyms. Notwithstanding, statistical tests between the slopes associated with normalized variants of each of these metrics were not significantly different, although there was a weak trend in the case of accuracy [for accuracy, $d = 0.030$, $SE = 0.019$, $t(418) = 1.6$, $p = .11$; for latency, $d = 5.96$, $SE = 34.2$, $t(418) = 0.24$, $p = .81$].

Next, we examined the robustness of these findings in multiple-regression analyses that also included several additional independent variables. Here, we report the results of analyses that included length (in letters), \log_{10} (SUBTL word frequency), orthographic Levenshtein distance, positional bigram frequency, number of phonemes, number of syllables, number of senses, number of verb interpretations, and number of noun interpretations as predictors. Simultaneous multiple regression was employed in which no interactions were allowed among these variables. We conducted separate analyses for β and U to avoid collinearity issues. Only the statistics related to the relative-frequency measurements are reported. In these analyses, neither variable predicted significant variance in either the accuracy or the latency data, although the effect of U was marginal when predicting latency (accuracy: β_f , $b = -0.0003$, $SE = -0.0002$, $p = .16$; β_u , $b = -0.0003$, $SE = 0.0002$, $p = .23$; U , $b = 0.0005$, $SE = 0.005$, $p = .93$; latency: β_f , $b = -0.06$, $SE = 0.19$, $p = .76$; β_u , $b = 0.4$, $SE = 0.3$,

⁷ The naming data were not used because semantic ambiguity effects are typically weak or nonexistent in those data (Borowsky & Masson, 1996).

$p = .18$; $U, b = -13, SE = 7, p = .08$). Finally, we repeated these regression analyses on a restricted set of the data that contained only the homonyms with fewer than 10 related senses associated with each of their meanings. This restriction limits the analyses to a subset of homonyms similar to those reported in the context of the analysis of the Armstrong and Plaut (2011) data, and which the aforementioned study suggested may reveal stronger effects of homonymy. There were 397 such homonyms in the full data set, and 110 in the data set for which U data were available. Only the analyses of β related to the set of words for which U data were available reached significance (accuracy: $\beta_f, b = -0.0005, SE = 0.0003, p = .11$; $\beta_u, b = -0.0009, SE = 0.0004, p = .04$; $U, b = -0.005, SE = 0.01, p = .66$; latency: $\beta_f, b = 0.05, SE = 0.3, p = .24$; $\beta_u, b = 0.95, SE = 0.46, p = .04$; $U, b = -10, SE = 11, p = .39$). The difference between the slopes associated with normalized variants of each of these metrics was marginal in the accuracy analysis [$d = 0.050, SE = 0.029; t(216) = 1.75, p = .08$] but was not significant in the latency analysis [$d = 24.0, SE = 30.2; t(216) = 0.79, p = .43$]. Consequently, although the simple-regression analyses weakly suggested that β was a significant predictor of a homonymy advantage (whereas U was not), the subsequent multiple-regression analyses suggest that this effect is at best extremely weak and is limited to homonyms with few related senses.

Discussion of the analyses In the analyses of the Armstrong and Plaut (2011) data, relative meaning frequency clearly altered the observed effects. Specifically, restricting the analyses to homonyms for which relative frequencies were balanced or including relative meaning frequency as a continuous variable generally allowed for the detection of a significant homonymy disadvantage. In contrast, the analyses of the data from the English Lexicon Project provide only very weak support for a general homonymy advantage in those data, and even that effect is restricted to when β , as opposed to U , is entered as a predictor. Of course, the weakness of many of these effects—which is attributable at least in part to not explicitly considering relative meaning frequency when selecting the experimental words—precludes drawing strong conclusions from these results. Nevertheless, these results support two tentative conclusions. First, relative meaning frequency should be considered when selecting experimental stimuli to maximize what appear to be, at best, weak effects of homonymy. Second, the presence of a homonymy advantage is particularly suspect, and it may be preferable to investigate the effects of homonymy in more difficult variants of the lexical decision task. However, additional experimental work using a range of tasks and carefully controlled sets of words will be needed to substantiate and generalize these conclusions. Automated means of selecting optimal stimuli (e.g., Armstrong et al., 2012) may be particularly helpful in these endeavors to maximize the magnitude of the apparently weak

effects of homonymy in visual lexical decision. Furthermore, the effects of homonymy should be examined in other tasks to establish the robustness of the present findings. Auditory lexical decision, in particular, may be worthy of further study, because some data suggest that the effects of homonymy are stronger in that task (Klepousniotou & Baum, 2007; Mirman et al., 2010; Rodd et al., 2002).

Conclusion

The present work outlines a method of assessing the relative meaning frequencies of each meaning of a homonym on the basis of dictionary definitions. We argued that this approach offers several theoretical and methodological advantages over standard norming approaches based on the classification of free associates. We collected normative data with the eDom software program for a large set of homonyms and presented evidence that supports the reliability and validity of this approach. Although additional work will be needed to better understand the unique aspects of meaning frequency that our method and other methods tap, as well as the effects of homonymy more generally, the present results motivate us to conclude that this norming method is to be preferred over standard methods based on free association.

Author note This research was supported by an NSERC Alexander Graham Bell Canada Graduate Scholarship to B.C.A. and by a Pennsylvania Department of Health grant to D.C.P. We thank the research assistants in the Tokowicz lab for help with data collection.

References

- Armstrong, B. C., & Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 273–278). Austin, TX: Cognitive Science Society.
- Armstrong, B. C., & Plaut, D. C. (2011). Inducing homonymy effects via stimulus quality and (not) nonword difficulty: Implications for models of semantic ambiguity and word recognition. In L. Carlson, C. Hölscher, & T. F. Shipley (Eds.), *Expanding the space of cognitive science: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 2223–2228). Austin, TX: Cognitive Science Society.
- Armstrong, B. C., Watson, C. E., & Plaut, D. C. (2012). SOS! An algorithm and software for the stochastic optimization of stimuli. *Behavior Research Methods*. doi:10.3758/s13428-011-0182-9
- Azuma, T., & Van Orden, G. (1997). Why SAFE is better than FAST: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language*, 36, 484–504.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items.

- Journal of Memory and Language*, 59, 390–412. doi:10.1016/j.jml.2007.12.005
- Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3(2), 12–28.
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133, 283–316. doi:10.1037/0096-3445.133.2.283
- Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445–459. doi:10.3758/BF03193014
- Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research*, 24, 57–65.
- Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 63–85. doi:10.1037/0278-7393.22.1.63
- Britton, B. K. (1978). Lexical ambiguity of words used in English text. *Behavior Research Methods & Instrumentation*, 10, 1–7. doi:10.3758/BF03205079
- 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, 977–990. doi:10.3758/BRM.41.4.977
- Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, 37, 65–70. doi:10.3758/BF03206399
- Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
- Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, 29, 181–200.
- Frisson, S., & Pickering, M. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1366–1383.
- Geis, M., & Winograd, E. (1974). Norms of semantic encoding variability for fifty homographs. *Bulletin of the Psychonomic Society*, 3, 429–431.
- Gilhooly, K. J., & Logie, R. H. (1980a). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12, 395–427. doi:10.3758/BF03201693
- Gilhooly, K. J., & Logie, R. H. (1980b). Meaning-dependent ratings of imagery, age of acquisition, familiarity, and concreteness for 387 ambiguous words. *Behavior Research Methods*, 12, 428–450.
- Gorfein, D., Viviani, J., & Leddo, J. (1982). Norms as a tool for the study of homography. *Memory & Cognition*, 10, 503–509.
- Hargreaves, I., Pexman, P., Pittman, D., & Goodyear, B. (2011). Ambiguous words recruit the left inferior frontal gyrus in absence of a behavioral effect. *Experimental Psychology*, 58, 19–30.
- Hino, Y., Kusunose, Y., & Lupker, S. (2010). The relatedness-of-meaning effect for ambiguous words in lexical-decision tasks: when does relatedness matter? *Canadian Journal of Experimental Psychology*, 64, 180–196.
- Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 1331–1356. doi:10.1037/0096-1523.22.6.1331
- Hino, Y., Pexman, P., & Lupker, S. (2006). Ambiguity and relatedness effects in semantic tasks: Are they due to semantic coding? *Journal of Memory and Language*, 55, 247–273.
- Hogaboam, T., & Perfetti, C. (1975). Lexical ambiguity and sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, 14, 265–274.
- Joordens, S., & Besner, D. (1994). When banking on meaning is not (yet) money in the bank: Explorations in connectionist modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 1051–1062. doi:10.1037/0278-7393.20.5.1051
- Kausler, D., & Kollasch, S. (1970). Word associations to homographs. *Journal of Verbal Learning and Verbal Behavior*, 9, 444–449.
- Kawamoto, A. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, 32, 474–516.
- Khanna, M. M., & Cortese, M. J. (2011). Age of acquisition estimates for 1,208 ambiguous and polysemous words. *Behavior Research Methods*, 43, 89–96. doi:10.3758/s13428-010-0027-y
- Klein, D., & Murphy, G. (2001). The representation of polysemous words. *Journal of Memory and Language*, 45, 259–282.
- Klein, D., & Murphy, G. (2002). Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language*, 47, 548–570.
- Klepousniotou, E., & Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics*, 20, 1–24.
- Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 1534–1543.
- Mirman, D., Strauss, T., Dixon, J., & Magnuson, J. (2010). Effect of representational distance between meanings on recognition of ambiguous spoken words. *Cognitive Science*, 34, 161–173.
- Neill, W., Hilliard, D., & Cooper, E. (1988). The detection of lexical ambiguity: Evidence for context-sensitive parallel access. *Journal of Memory and Language*, 27, 279–287.
- Nelson, D. L., McEvoy, C. L., Walling, J. R., & Wheeler, J. W., Jr. (1980). The University of South Florida homograph norms. *Behavior Research Methods & Instrumentation*, 12, 16–37. doi:10.3758/BF03208320
- Parks, R., Ray, J., & Bland, S. (1998). *Wordsmyth English dictionary–thesaurus* (vol. 1). Retrieved September 2008 from wordsmyth.net.
- Pickering, M., & Frisson, S. (2001). Processing ambiguous verbs: Evidence from eye movements. *Learning and Memory*, 27, 556–573.
- Plaut, D. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language & Cognitive Processes*, 12, 765–805.
- Pylkkänen, L., Llinás, R., & Murphy, G. (2006). The representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience*, 18, 97–109.
- Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 46, 245–266.
- Rodd, J., Gaskell, M., & Marslen-Wilson, W. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, 28, 89–104.
- Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, 9, 487–494. doi:10.1016/S0022-5371(70)80091-3
- Schvaneveldt, R. W., Meyer, D. E., & Becker, C. A. (1976). Lexical ambiguity, semantic context, and visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 243–256. doi:10.1037/0096-1523.2.2.243
- Seidenberg, M., Tanenhaus, M., Leiman, J., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology*, 14, 489–537.
- Simpson, G. (1984). Lexical ambiguity and its role in models of word recognition. *Psychological Bulletin*, 96, 316–340.
- Simpson, G. (1994). Context and the processing of ambiguous words. In Gernsbacher Morton Ann (Ed.), *Handbook of psycholinguistics* (pp. 359–374). San Diego, CA: Academic Press.

- Swinney, D. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*, 645–659.
- Tokowicz, N., & Kroll, J. F. (2007). Number of meanings and concreteness: Consequences of ambiguity within and across languages. *Language & Cognitive Processes*, *22*, 727–779.
- Twilley, L., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition*, *22*, 111–126.
- Warren, R., Bresnick, J., & Green, J. (1977). Definitional dominance distributions for 20 English homographs. *Bulletin of the Psychonomic Society*, *10*, 229–231.
- Williams, J. (1992). Processing polysemous words in context: Evidence for interrelated meanings. *Journal of Psycholinguistic Research*, *21*, 193–218.
- Wollen, K., Cox, S., Coahran, M., Shea, D., & Kirby, R. (1980). Frequency of occurrence and concreteness ratings of homograph meanings. *Behavior Research Methods*, *12*, 8–15.
- Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971–979. doi:[10.3758/PBR.15.5.971](https://doi.org/10.3758/PBR.15.5.971)
- Yates, J. (1978). Priming dominant and unusual senses of ambiguous words. *Memory & Cognition*, *6*, 636–643.