A Comparison of Homonym Meaning Frequency Estimates Derived from Movie and Television Subtitles, Free Association, and Explicit Ratings

Caitlin A. Rice

Department of Psychology

University of Pittsburgh & Center for the Neural Basis of Cognition

Barend Beekhuizen

Department of Computer Science

University of Toronto

Vladimir Dubrovsky

Department of Computer Science

University of Toronto

Suzanne Stevenson

Department of Computer Science

University of Toronto

Blair C. Armstrong

Department of Psychology and Center for French & Linguistics at Scarborough

University of Toronto

Basque Center on Cognition, Brain, and Language

Author Note

Correspondence regarding this article should be sent to Caitlin Rice, Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara St., Pittsburgh, PA 15260, USA, e-mail: car120@pitt.edu. We thank our large group of volunteer raters for assistance in rating the free association and movie subtitle data. CAR was supported by a Women in Cognitive Science International Networking Award. BCA was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (2017-06310). BB, VD, and SS were supported by NSERC Discovery Grants (227787 and 2017-06506). VD was also supported by an NSERC Undergraduate Student Research Award (USRA).

Abstract

Most words are ambiguous, with interpretation dependent on context. Advancing theories of ambiguity resolution is important for any general theory of language processing, and for resolving inconsistencies in observed ambiguity effects across experimental tasks. Focusing on homonyms (words such as *bank* with unrelated meanings EDGE OF A RIVER vs. FINANCIAL INSTITUTION), the present work advances theories and methods for estimating the relative frequency of their meanings, a factor that shapes observed ambiguity effects. We develop a new method for estimating meaning frequency based on the meaning of a homonym evoked in lines of movie and television subtitles according to human raters. We also replicate and extend a measure of meaning frequency derived from the classification of free associates. We evaluate the internal consistency of these measures, compare them to published estimates based on explicit ratings of each meaning's frequency, and compare each set of norms in predicting performance in lexical and semantic decision mega-studies. All measures have high internal consistency and show agreement, but each is also associated with unique variance, which may be explained by integrating cognitive theories of memory with the demands of different experimental methodologies. To derive frequency estimates, we collected manual classifications of 533 homonyms over 50 000 lines of subtitles, and of 357 homonyms across over 5000 homonym–associate pairs. This database—publicly available at: `www.blairarmstrong.net/homonymnorms/`—constitutes a novel resource for computational cognitive modeling and computational linguistics, and we offer suggestions around good practices for its use in training and testing models on labeled data.

Keywords: Semantic ambiguity; homonyms; meaning frequency; homonym norming methods and data; movie subtitles; free association; homonym meaning annotations

A Comparison of Homonym Meaning Frequency Estimates Derived from Movie and Television Subtitles, Free Association, and Explicit Ratings

Introduction

One striking property of natural language is that the same word is typically associated with different interpretations in different contexts. The ability to disambiguate a word's interpretation based on context is therefore an essential component of any theory of language comprehension. The importance of developing an account of how the various interpretations of a word are represented, as well as how different interpretations can be selectively activated, is highlighted by the extensive empirical and computational literatures devoted to these interrelated issues over the past several decades (Armstrong, Tokowicz, & Plaut, 2012; Armstrong, Zugarramurdi, Cabana, Valle Lisboa, & Plaut, 2015; Armstrong & Plaut, 2016; Gernsbacher, 1984; Hino, Pexman, & Lupker, 2006; Hino, Kusunose, & Lupker, 2010; Kawamoto, 1993; Klepousniotou, 2002; Klepousniotou, Titone, & Romero, 2008; Klepousniotou, Pike, Steinhauer, & Gracco, 2012; Piercey & Joordens, 2000; Rodd, Gaskell, & Marslen-Wilson, 2002, 2004; Swinney & Hakes, 1976; Swinney, 1979; Tabossi, 1988; Twilley, Dixon, Taylor, & Clark, 1994; Williams, 1992). Collectively, these studies have revealed a range of complex and sometimes apparently contradictory effects of ambiguity, making great strides towards understanding the critical facets of word representation that shape performance.

One factor that was under-appreciated in much prior work is not only how many interpretations a word is associated with, but whether those interpretations are related to one another or not (Rodd et al., 2002). By taking relatedness of interpretation into account, a number of studies have revealed different empirical effects for words that are homonymous, and are associated with multiple unrelated meanings (e.g., *bank* refers to the EDGE OF A RIVER and to a FINANCIAL INSTITUTION), versus words that are polysemous, and are associated with multiple related senses (e.g., *paper* refers to an ACADEMIC ARTICLE and a WHITE SHEET) (Armstrong & Plaut, 2016; Rodd et al., 2002; Hino et al., 2006, 2010; Klepousniotou et al.,

2008). For example, a common pattern is that homonyms tend, at least numerically, to incur a processing disadvantage when compared to (relatively) unambiguous control words (e.g., *chalk*), while in contrast, polysemes tend to show a processing *advantage* relative to unambiguous words. The particular patterns of statistical significance related to this homonymy disadvantage and polysemy advantage do, however, vary considerably across tasks, and there are theoretically important exceptions, for example, showing a processing advantage for homonyms (e.g., Hino et al., 2006, 2010).

What might explain the variability in these results? One apparently contributing factor (among many, such as the decision system, Hino et al., 2006) is the relative frequencies of the different interpretations of a word. This influence was mentioned in some of the early semantic ambiguity literature (e.g., Swinney, 1979; Swinney & Hakes, 1976), but has received only moderate attention in mainstream semantic ambiguity research, particularly in tasks that probe the processing of ambiguous words in the absence of biasing context (for discussion, see Armstrong & Plaut, 2016). For example, it may take more time to strongly activate one meaning of a homonym that has meanings of approximately equal frequency (e.g., *compound*, which refers to a MIXTURE and an ENCLOSURE) because the meanings compete with one another, initially inhibiting either from becoming strongly activated. In contrast, it may take less time to activate the meaning of a less balanced homonym (such as *bank*) because its dominant meaning (in this case, FINANCIAL INSTITUTION) can rapidly inhibit the subordinate meaning. Indeed, these effects bear a close resemblance to the classic "rich get richer" effects which are a hallmark of many general information processing frameworks (e.g., McClelland & Rumelhart, 1981). Moreover, these competitive dynamics should be more pronounced for homonyms relative to polysemes: In contrast to the unrelated meanings of homonyms, the overlapping interpretations of a polyseme would all contribute to the activation of shared meaning features, engendering less competition. According to this view, taking into account the relative frequency of homonym meanings should be of particular importance to understanding different patterns of effects

associated with those items, a proposal that has garnered at least some empirical support (e.g., Armstrong & Plaut, 2016; Frazier & Rayner, 1990; Klepousniotou & Baum, 2007; Williams, 1992).

Deriving good estimates of the relative meaning frequencies of a homonym is therefore a critical question from both theoretical and methodological perspectives, and understanding how different methods yield similar (or dissimilar) results is the primary focus of this article. Past research has pursued various methods for deriving estimates of meaning frequency from experimental tasks such as sentence generation and free association (for norms and a review, see Twilley et al., 1994), and recent work has also explored the efficacy of collecting explicit meaning frequency ratings (Armstrong, Tokowicz, & Plaut, 2012). At face value, if each of these tasks were tapping into the same underlying representation of meaning frequency in a similar fashion, the derived measures should all converge on similar results. Yet, this research has revealed just how challenging it is to determine robust and predictive estimates of meaning frequencies. For example, measures of relative meaning frequency derived from explicit ratings and from a free association task have been found to correlate only weakly (for the Twilley et al. and Armstrong et al. data, $r = .27$; Armstrong, Tokowicz, & Plaut, 2012). Similarly, these different predictors vary considerably in their external validity – that is, in terms of their ability to predict performance in tasks such as lexical decision (cf. the findings in Armstrong, Tokowicz, & Plaut, 2012). Furthermore, although the collection of all such empirical measures entails a relatively high cost in terms of time and laboratory resources, the different methods also vary in the amount of data needed to elicit reliable estimates of meaning frequency, with tasks based on free association requiring one or two orders of magnitude more participants to achieve stable norms than explicit ratings of meaning frequency. Understanding how these different measures align therefore remains an important and pressing issue for further progress in semantic ambiguity research.

Advancing methods for the norming of relative meaning frequency also offers an excellent opportunity for synergy between the psycholinguistic and computational linguistic communities.

With the exception of explicit meaning frequency estimates, the other tasks used to measure semantic ambiguity involve the rating of some experimentally generated data such as a free association or a sentence completion to see which meaning of a homonym was evoked in that context. In computational linguistics, the similar task of word sense[1] disambiguation — automatically identifying which meaning of an ambiguous word has been evoked in a particular context — has been an important and long-studied theme (*Word Sense Disambiguation: Algorithms and Applications*, n.d.; Bartunov, Kondrashkin, Osokin, & Vetrov, n.d.; Lefever & Hoste, 2010; Li & Jurafsky, 2015). Recent work in this field has also focused on developing unsupervised methods for determining relative meaning frequencies, or sense distributions for ambiguous words, including but not limited to homonyms (e.g., Bennett, Baldwin, Lau, McCarthy, & Bond, 2016; Lau, Cook, McCarthy, Gella, & Baldwin, 2014). Bringing these psycholinguistic and computational linguistic strands of work together, it may be possible to partially or fully automate the estimation of homonym meaning frequencies both in natural language and in the behavioural responses derived from empirical studies of semantic ambiguity.

This aim should be facilitated by the availability of our new set of labeled data, in which multiple human raters have indicated the correct meaning for hundreds of homonyms in context. Although, as we discuss in detail in the discussion, we are far from the first to generate this general type of "annotated" dataset, we view our work as offering complementary coverage to many prior efforts in computational linguistics for several reasons: (1) We focus on conversational natural language corpora (our norms based on movie and television subtitles) and experimental data (our norms based on a free association task) that are well-established in the psycholinguistic literature and well suited for merging cognitive theory with computational linguistic theory. This is important given that past research has shown that the genre of a text is relevant to disambiguating words (Gale, Church, & Yarowsky, 1992). (2) Our classifications are

[1] In the computational linguistics literature, the term "sense" in "word sense disambiguation" is used differently than in the psycholinguistic literature where "sense" is typically used to refer to a related interpretation of a polyseme. In computational linguistics, this term would also cover the disambiguation of homonyms.

based on the definitions offered in the Wordsmyth dictionary, which also has a long history of use for deriving measures of homonymy and polysemy in the psycholinguistic literature. (Cf. many other efforts based on WordNet, which does not delineate between unrelated meanings and related senses; for an overview of work based on WordNet, see Petrolito & Bond, 2014.) (3) By collecting confidence of ratings on a broad selection of homonyms in natural contexts, the annotations should also be informative on which contexts are more or less helpful for people in the disambiguation task, which is relevant for developing models that process language the same way that humans do. We return to this issue in the discussion.

With the set of goals outlined in the previous paragraphs in mind, we embarked upon the empirical studies of relative meaning frequency that we describe next. As a guiding assumption, we reasoned that there should be a broad alignment between how often people encounter a homonym in each of its meanings in their linguistic experience, and how these meanings are relatively activated in an experimental task (whether an indirect experimental assessment such as a free association norming task, or an explicit meaning frequency rating task). We therefore collected two extensive sets of ratings, one of the meanings a given homonym elicits in the natural language context of lines of movie and television subtitles, and another of the meanings that appear to have been activated to generate particular associates in a free association task. These two datasets were then subject to analyses that evaluated the reliability of the ratings, and, after converting the rated data into estimates of meaning frequency, were compared to existing meaning frequency estimates (Armstrong, Tokowicz, & Plaut, 2012; Twilley et al., 1994).

To preview our results, there was general agreement but also a number of important and potentially theoretically informative differences between the norms collected using each method. These results therefore have both practical implications for how to measure and evaluate estimates of meaning frequency, as well as theoretical implications regarding how knowledge of meaning frequency is stored and retrieved in different contexts. In service of facilitating related and complementary research into these issues, we have made all of our rated data and meaning

frequency estimates available for download at:
www.blairarmstrong.net/homonymnorms/.

Estimating homonym meaning frequency from free association norms

The first set of meaning frequency estimates that we created were derived from the Nelson, McEvoy, and Schreiber (2004) free association norms. Our motivation to estimate homonym meaning frequencies from free association norms was twofold. First, this is one of the most popular and long-standing methods for assessing homonym meaning frequency (Nelson, McEvoy, Walling, & Wheeler, 1980; Twilley et al., 1994). Following the standard practice for collecting free associates, participants are presented with a word (the cue) and respond to the cue with the first word that comes to mind (the associate). When homonyms are used as cues, raters can classify which meaning of the homonym is most closely related to the associate. The total number of associates linked with each meaning can be used to derive estimates of relative meaning frequency, based on the assumption that participants generate associates related to each meaning as a function of the meaning's frequency. Our norming process here similarly began with classifying each associate based on the original normed data in Nelson et al. (2004).

Second, as noted earlier, previous work has found a surprisingly low correlation between one set of meaning frequencies estimated from free association norms (Twilley et al., 1994) and explicit ratings of meaning frequency (Armstrong, Tokowicz, & Plaut, 2012). Here, we evaluate the robustness of this finding in a dataset that contains approximately 50% more homonyms. Moreover, in contrast to the free association task of Twilley et al. (1994), the Nelson et al. (2004) data also includes non-homonyms as cues; this enables us to evaluate if the previous findings were particular to that set of homonyms and specific set of methods. Proceeding from our simple assumption regarding how meanings are represented and how these representations are tapped in either indirect tasks (e.g., classification of free associates) or direct tasks (e.g., explicit meaning frequency estimates), we predict a higher correlation between explicit ratings and the subtitles

measure than between explicit ratings and the free association norms. This is because explicit ratings and free association both aim to hone in on frequency effects, whereas indirect measures based on free association may be less sensitive to meaning frequency because they are influenced by other cue-associate relationships. For example, these relationships include whether two words rhyme, can combine to make a new word (e.g., wrist-watch), are category coordinates (e.g., robin-sparrow), and so on (Armstrong, Tokowicz, & Plaut, 2012).

Estimating homonym meaning frequency from movie and television subtitles

How do individuals learn the frequency with which a homonym is used to denote a particular meaning? According to recent empirical work, much in the same way that they learn that some words are more frequent than other words: through exposure to natural language contexts (Rodd et al., 2016). Previous work has shown that movie and television subtitle corpora can be used to derive measures of word frequency in natural language contexts that are excellent predictors of performance in a range of psycholinguistic tasks (Brysbaert & New, 2009). We reasoned that a similar approach could be used to derive estimates of meaning frequency (for an earlier related effort, see Lorge, 1937). To this end, we extracted random samples of text that included our target homonyms, and raters then indicated which meaning of the homonym was evoked in each sample. Again, if our simple assumption is correct regarding how meaning frequencies are learned, internally represented, and retrieved in different tasks, we expect this new measure to correlate well with our other estimates, and particularly the explicit meaning frequency norms (given that a free associate may have been generated for reasons other than meaning frequency alone). The classified text generated as an intermediate step in calculating meaning frequencies is also likely to be of particular interest to computational linguists that require large labeled natural language datasets to train and test models of meaning disambiguation.

Comparison to existing meaning frequency norms

We compare our two new measures of meaning frequency with two previously established sets of meaning frequency estimates. The first are the original eDom explicit estimates of homonym relative meaning frequency by Armstrong, Tokowicz, and Plaut (2012). There, participants were presented with each dictionary definition of a homonym and asked to estimate the frequency with which they thought the word was used to denote each of those meanings in everyday use. (Participants could also add a definition they thought was missing.) Critical for our purposes, participants very rarely generated such additional definitions although they did so reliably when there was a high frequency interpretation not listed in the dictionary. This suggests that dictionary definitions—although not perfect in their coverage of word meanings—provide more than sufficient coverage of the meanings of the vast majority of words to serve as the basis for classification in the present experimental tasks.

The second comparison set are estimates of meaning uncertainty derived from the classification of free associates reported by Twilley et al. (1994), which we refer to as Twilley *U*. Twilley and colleagues' procedure for deriving a measure of meaning uncertainty began as outlined above, with classifying associates according to the related meaning of the homonym cue. They then used these meaning frequencies to derive a measure of uncertainty, *U* (based on Shannon Entropy), which the authors proposed as the relevant measure to capture meaning distribution. Comparisons to these two additional sets of norms – eDom *biggest* and Twilley *U* – are particularly interesting because they will allow us to evaluate the robustness of the surprisingly weak relationship between explicit meaning frequency estimates and indirect estimates based on the classification of free associates.

Methods

Because the methods were very similar for the classification of free associates and for the film and television subtitles, we report all of the methods for both tasks before proceeding to the

results from both datasets. We begin by describing the methods for the free associate classification task; for the subtitles classification task, we only report the ways in which that task differed from the first task.

Deriving Homonym Meaning Frequencies from Free Association Norms (FAN)

Raters. Raters consisted of nine undergraduate students from the University of Toronto who were provided with training on how to classify associates according to the meanings listed in the dictionary, and who rated the first several word associates together under the supervision of one of the authors. After that point, they rated all words independently of one another. All raters were either enrolled as undergraduate students at the University of Toronto Scarborough or had recently graduated from an undergraduate program and were highly proficient speakers of English (mean age = 21, range = 18–24; 8 Female; 7 Native English speakers).

Stimuli. The free association stimuli examined by the raters were sampled from the South Florida Free Association Norms (Nelson et al., 2004), downloaded from `http://w3.usf.edu/FreeAssociation/Intro.html`. The use of this well-established set of norms was particularly appealing for our purposes both because of its large size and because many of the cue words were selected from prior free association studies and were also homonyms (Nelson et al., 1980).

To enable comparisons between identical homonyms across multiple datasets, we attempted to extract data from the free association norms for all homonyms that were also reported in the eDom meaning frequency norms (Armstrong, Tokowicz, & Plaut, 2012). The original eDom norms contained 544 homonyms, selected to satisfy standard criteria for use in psycholinguistic experiments, such as falling within typical ranges in terms of length and frequency (for a full description and rationale, see Armstrong, Tokowicz, & Plaut, 2012). In total, 364 of these 544 eDom homonyms were also present in the South Florida Norms. For reference, this was over 150 more overlapping homonyms than were found between eDom and the meaning frequency norms

based on free association reported by Twilley et al. (1994). For these 364 homonyms, we extracted all unique associates produced for each homonym, as well as the number of participants that produced each associate in the original study. To improve the reliability of the original norms, the data published by Nelson et al. (2004) only include associates produced by two or more participants. A total of 5176 associates were extracted from the database, with a mean of 14 associates for each homonym (SD = 5, range = 2–29).

Procedure. In preparation for the rating task, the definitions for each homonym were retrieved from the Wordsmyth Dictionary (Parks, Ray, & Bland, 1998), using the same parse as in the creation of the eDom norms. The Wordsmyth dictionary has been used in a number of prior studies of semantic ambiguity, and meaning and sense estimates derived from this dictionary have been used to estimate the number and relatedness of a word's interpretations (e.g., Armstrong & Plaut, 2016; Armstrong, Tokowicz, & Plaut, 2012; Rodd et al., 2002). In this dictionary, separate entries are used for each unrelated meaning of a word, whereas sub-entries within each entry are used to denote related senses. The raters were then given these definitions to guide their classification of homonym free associates according to the different entries. Because the focus of this project is on homonymy, as opposed to polysemy, no attempt was made to distinguish among sub-entries (which list related senses of a word). However, to gain some preliminary insights related to polysemy, we do include number of senses as a covariate in some of the analyses reported in the results section.

It is worth noting that although past literature clearly supports dictionary definitions as a valid basis for identifying and classifying word meanings, it is not the only method that has been explored to this end. For example, considerable work has used subjective ratings to identify word meanings by asking different groups of participants generate and rate the relative frequency of word meanings (e.g., Hino et al., 2006; Pexman, Lupker, & Hino, 2002). Other work has examined the utility of delineating word meanings based on classification schemes outlined in the theoretical linguistics literature (Klepousniotou & Baum, 2007; Klepousniotou et al., 2012). Each

of these approaches has its own strengths and a full discussion and determination of which approach is "best" is outside the scope of the present work. For our purposes, we opted to use the dictionary definition approach for two reasons. First, because it has been documented as able to successfully tap into human representations of word meaning (regardless of whether it is the best method to do so), Second, the availability of dictionaries in diverse languages makes this approach less costly in terms of collection time and/or the need for experts in linguistic classification.

Similar to past norming procedures (e.g., Twilley et al., 1994), for each associate, the raters had to determine which of the definitions of the homonym (if any) best fit with the associate. They entered their responses as a numerical value based on the ranked order in which that definition appeared in the dictionary (e.g., entry 1, entry 2, etc.), or if none of the meanings fit, flagged that associate with a dash (-). Supplementing standard classification methods, the raters also provided a coarse coding for the confidence of their responses: either they were certain that their classification was correct (in which case, they simply entered a "1" for entry 1, a "2" for entry 2, and so on), or they were uncertain but thought their classification was the best guess given the available options (in which case they prefaced their numerical value with a "?", e.g., "?1", "?2"). Raters could also provide a comment related to their classification when appropriate. As discussed in the results section, these comments often help explain the link between the cue and the associate when a dictionary definition does not fit. As a couple of examples, some cue-associate relationships were attributable to a colloquial meaning of the cue not included in the dictionary (such as the case of *rip* referring to an insult), or to the cue and associate rhyming with one another.

To obtain a measure of interrater reliability, three independent ratings were collected from different raters for each associate. A rater classified all of the associates for a given homonym before moving on to another randomly selected homonym. The task was completed in a large number of small batches, with raters typically working on the task for a half hour several times per week, with the only restriction being that the full set of associates for a given homonym

should always be classified within the same rating session. Typically, raters completed approximately 300 ratings per hour. In total, this entailed over 15 000 separate classifications and over 50 hours of rater time. On average, each rater completed 1941 classifications (SD = 1758, range = 104–4955, inter-quartile range = 487–2988).

Deriving Homonym Meaning Frequencies from Video Subtitles (SUBTL)

The second norming task was analogous to the first, except that instead of classifying which homonym meaning was evoked by a particular free associate, raters classified which homonym meaning was evoked by a line of dialog extracted from a corpus of movie and television subtitles. Except as described below, all methods for the second task were identical to those used in the first task.

Raters. The set of raters was expanded to include 18 raters, 7 of whom had previously contributed to the classification of free associates. All but two of the new raters were University of Toronto Scarborough undergraduate students; the remaining raters were an undergraduate at the University of Pittsburgh, and the first author, who only contributed to the re-rating of 413 cases for which one rater appears to have flipped the code for dictionary entry 1 and 2 (raters' mean age = 21, range = 18–29; 14 Female; 15 Native English speakers).

Stimuli. Stimuli consisted of individual lines of subtitles extracted from the same corpus of movie and television subtitles used to derive the SUBTL word frequency estimates (Brysbaert & New, 2009), downloaded from `http://subtlexus.lexique.org/corpus/Subtlex%20US.rar`. For copyright reasons, each line appears in a random order in the corpus. Therefore, participants are effectively rating each subtitle line without access to broader information about the context in which the line occurs.

Similar to the procedure used for the free association norms, we extracted all lines in which the lower case form of one of the 544 homonyms in the eDom norms appeared (we focused on the

lower case condition to avoid cases where a word was used in a proper noun form). Each homonym occurred on average 641 times in the SUBTL corpus (SD = 907, min = 1, max = 4695; IQR = 103–690; all values are for lower-case use). To keep the rating task tractable, we then extracted a random sample of 100 lines of subtitles for each homonym. For the few homonyms where there were less than 100 lines, we included all available lines in the rating task. This yielded a total of 49 156 lines to rate across 544 homonyms, for a mean of 91 lines per homonym (range: 1–100; however, samples at the low end of this range were very rare; IQR = 100—100, only 24% of the homonyms had less than 100 lines, only 6% had less than 50 lines, and <1% had less than 10 lines).

One potential concern with the use of this database, particularly for homonyms with a low frequency, is that these homonyms may appear in one or only a small number of different films or television programs and the themes of those programs could bias the meaning frequency estimates. Further inspection of the distribution of homonyms across the different movie and television programs (hereafter, contexts) alleviated this concern, as is illustrated in Figure 1. This figure plots the log transformed number of unique contexts as a function of log-transformed word frequency for all of the homonyms (context counts were also taken from Brysbaert & New, 2009). It shows a very strong positive relationship (r = .97) between word frequency and number of contexts, and that the vast majority of homonyms are occurring in hundreds of contexts (mean = 414, SD = 522).

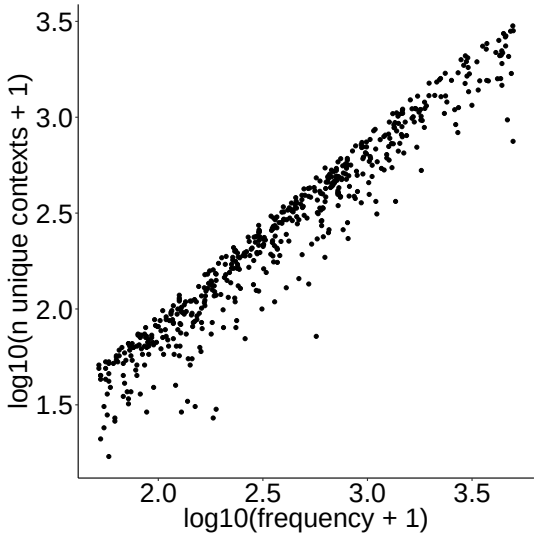

Figure 1. Scatterplot of number of unique contexts in which the word occurred in the SUBTL corpus as a function of word frequency.

Procedure. The rating task was identical to that used for the free associates, except that due to the large size of the dataset, it was not feasible to have three raters classify every line of subtitles. Instead, we collected two independent ratings for every subtitle line in the dataset, totalling just under 100000 ratings. Although the task was to read lines of subtitles instead of single associates, we found that raters continued to complete approximately 300 ratings per hour. Thus, collecting these new norms required approximately 335 hours of work. The mean number of classifications completed by each rater was 5464 (SD = 7929, range = 68–26895, IQR = 1048–5464).

Data Cleaning

While completing the norming task, we discovered that eight of the original Wordsmyth definitions had not been extracted from the dictionary correctly, and so we removed the affected homonyms (not all of which were present in every corpus) from all datasets and analyses (*corrupt, costume, cottage, cotton, couch, cough, counsel, and rack*). Similarly, the definitions for the meanings of two words (*lit*, *rung*) were extremely impoverished and focused on how one

definition related to a past participle whose root form was defined under the main dictionary entry, and were removed from all subsequent analyses. Finally, ten of the original eDom words were homographs (*bow, bowman, colon, content, lead, mare, present, wind, and wound*), and so were removed from all datasets and analyses. After these exclusions, 534 homonyms remained in the eDom dataset, 357 homonyms remained in the FAN dataset, 533 homonyms remained in the SUBTL dataset, and 205 homonyms remained in the Twilley dataset.

FAN. Before analyzing the FAN data, we first inspected all of the data for which participants either could not link an associate with any of a homonym's definitions and/or for which they had made a comment that flagged a particular associate as unusual in some respect. The comments revealed that many of these associations could readily be explained either by meanings of words not included in the dictionary (e.g., proper nouns, as in *dove–soap*; slang meanings as in *hip–in*), or by associations that are shaped by phonological, as opposed to purely semantic, relationships (e.g., *sage–rage*). The presence of such data was to be expected in a completely free association task, and has been one of the reasons that meaning frequencies based on free association norms have been criticized in the past (Armstrong, Tokowicz, & Plaut, 2012). However, to date, the number of associates linked to representations other than those of the dictionary meanings of a homonym had not been quantified, so it was somewhat surprising to us that such cases represented less than 2% of the data (93 associates). This suggests that free association norm tasks are almost exclusively driven by semantic associations related to the meanings of each word and that dictionary definitions come close to providing exhaustive coverage of the meanings of words evoked in this context. So as to avoid additional noise or influence from this small number of unusual cases, they were removed from all computations of relative meaning frequency, although these data were left in for our initial plot of internal validity reported in the next section to illustrate the small size of this subset.

SUBTL. As before, following completion of the ratings, one of the authors manually inspected cases for which raters listed a comment and found that a large number of these

comments related to slang interpretations of a homonym not covered in the dictionary (e.g., the JAIL meaning of *slam*: "The reason he got locked in the *slam* in the first place ... was for sticking up a gas station to cover you guys"). The SUBTL database was intentionally used because it reflects the more naturalistic use of word meanings beyond how those meanings are defined by lexicographers, so the presence of such colloquial uses is to be expected. Raters left comments for 2% of the lines of subtitles (1037 lines) and 45% of the commented lines indicated that the homonym referred to a meaning other than one present in the dictionary definitions. As for the FAN data, cases for which one or both raters indicated a line of dialog captured a meaning that was not captured by Wordsmyth were removed from all computations of meaning frequency but were left in for our initial plot of internal validity.

Results

In the following section, we perform analyses over our two new datasets of FAN and SUBTL, comparing them to eDom and Twilley *U*. Note that whereas the eDom data and SUBTL data both contain data from over 500 homonyms, data from fewer homonyms were available for the FAN and *U* datasets (n = 357 and 205, respectively). Because of these substantial differences in sample size and potential generalizability of the results, in the main text we report the results of analyses using the full set of data available for each individual measure, but in addition to these results we also offer, in the Appendix, parallel analyses using only the intersecting data available for multiple measures.

Dataset	Number of words	Type of words	Task	Total number of raters	Number of ratings per word**
eDom	533	Primarily homonyms; a few homographs	Explicit ratings of dictionary definitions	64	16
FAN	357	Homonyms	Classification of free associates	9	3
SUBTL	533	Homonyms	Ratings of corpus usage	18	2
Twilley	205	Homonyms and Homographs	Classification of free associates	139*	192

Table 1

*Comparison of datasets. In the original eDom article, data for 544 homonyms were reported but 11 were dropped in the present analyses, as detailed in the main text. In the original Twilley article, data from 566 homographs and homonyms were reported; our analyses focus on the 205 homonyms that overlapped with the eDom norms. * The total number of raters for the full set of 566 homographs was 384; to facilitate comparisons with the other datasets, we scaled this number relative to the 205 out of 566 words analyzed here. **For FAN and SUBTL this number refers to the exact number of raters per word; for eDom and Twilley this number reflects the average number of ratings per word*

Internal validity

FAN. Previous studies of free association norms have reported that human raters have some difficulty classifying associates as belonging to particular dictionary definitions. For example, in the original Twilley norms (Twilley et al., 1994), raters agreed on how to match an associate with a definition only 65% to 75% of the time. However, for reasons that are not readily apparent, our data paint a much different picture of the reliability of free associate classifications. Specifically, when we examined inter-rater agreement in the FAN norms, we found that rater agreement was high, with 91% of the associates being given the same classification by all raters and all raters expressing high confidence in their classifications. Hereafter, we refer to this highly confident and consistent agreement as "certain agreement". This agreement level increases to 93% if we also include cases where raters agreed but expressed reduced confidence in their ratings.

Figure 2 provides additional insight into the distribution of classifications and amount of (dis)agreement across raters. The first two bars of this plot (labeled CA1 and CA2) represent the percentage of responses that correspond to certain agreement that an associate is linked to the first or second definition, respectively, in the Wordsmyth dictionary. These bars indicate that 76% of the time, all raters confidently agreed that the first meaning was the basis for generating the associate, and 14% of the time, the second meaning was the basis for generating an associate (90% of the data). The third bar (labeled CD) plots the percentage of cases for which at least two of the three raters were confident of their ratings but disagreed on which meaning of a homonym was used; it shows that such cases were rare (5% of the data). Manual inspection indicated that for some of these responses, there are potential links from both meanings to the associate (e.g., the associate *match* produced for the cue *box* could denote a box of FLAMMABLE MATERIAL or an ATHLETIC COMPETITION). The final bar (labeled Other) represents the sum of all other cases. These cases broke down into those in which raters agreed on the meaning evoked for a particular associate but expressed low confidence in their rating (2% of the data); cases where multiple raters disagreed but expressed low confidence in their judgments, ($< 1\%$ of the data);

cases where all raters agreed that no definition in Wordsmyth captured the represented meaning (2% of the data); and cases where raters classified an associate as linked to a third meaning of a homonym ($< 1\%$ of data).

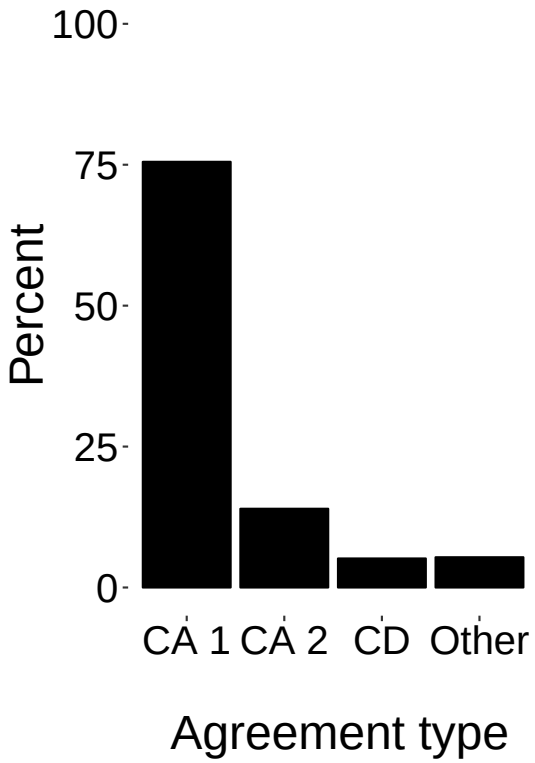

Figure 2. Distribution of ratings as a function of rater agreement in the FAN data. CA1, CA2: Certain Agreement that the associate was linked to meaning 1 or meaning 2, respectively. CD: Raters were certain about their classifications but their classifications disagreed. Other: All other cases.

Given the high amount of "certain agreement" among the raters, the CA1 and CA2 bars in Figure 2 can be interpreted as plotting the frequency with which the dominant and subordinate meanings of a homonym were evoked in the free association task. To make a strong claim in this regard, however, we needed to also rule out a simple and less interesting explanation for our results: raters simply have a bias to indicate that the first dictionary meaning underlies the association, for instance, because this is the first definition that they saw. If this were the case, such a bias to choose the first rating might artificially inflate the levels of agreement such that

what appears to be high agreement may actually be due to chance once bias was factored out.

To evaluate this possibility, we conducted a Monte Carlo (MC) simulation to assess the likelihood that the observed level of agreement would be expected due to chance while controlling for rater bias. (See the Appendix for motivation for this approach over standard parametric tests of inter-rater reliability, as well as additional details regarding the method and results.) We found the actual level of agreement was vastly higher than that of the chance distribution, and none of the simulations generated a level of agreement larger than that observed empirically (i.e., the probability that we would have observed our level of agreement by chance is zero in the simulations). We conclude that our high levels of inter-rater agreement are clearly not attributable to a simple bias for raters to produce classifications that favour the first meaning, or due to chance alone.

Taken together, these data indicate that, by and large, raters were highly confident in their responses, in contrast to the previous report by Twilley et al. (1994). Most associates were also linked with one of the first two dictionary definitions of a homonym and rarely to a third meaning if a homonym had one. This finding replicates a similar observation from the original eDom norms and indicates that the dictionary provides relatively exhaustive coverage of the meanings of words, even if some of those meanings are used very infrequently. For this reason, all subsequent results only plot data related either to only meaning 1 or to meaning 1 and meaning 2, which represent the vast majority of both the FAN data, as well as the SUBTL data, which are described next.

SUBTL. Inter-rater reliability was assessed in a similar manner for the SUBTL data. The overall pattern of agreement was extremely similar to that observed for the FAN norms, as illustrated in Figure 3. Of particular note was that the two raters produced the same highly confident classifications linking the meaning of a homonym in a line of subtitles to the first and second definitions of a homonym, which represented 73% and 17% of the data, respectively, thus covering 90% of cases (if both uncertain and certain agreement between raters were counted, this

total rises to 92%; cf. 76%, 14%, 90%, and 92% for the analogous measures derived from the FAN data). Similar low rates of confident, disagreeing classifications between the two raters were also observed (e.g., raters confidently disagreed about whether "I'm going to *cuff* you" should be classified as an example of the HANDCUFF or the HIT meaning of *cuff*), as well as of "other" types of responses (a subset of the "other" data represents classifications pertaining to an additional dictionary definition other than the first two definitions, which occurred less than 2% of the time, e.g., "You have to *buck* up and do it"). Once again, we conducted an MC simulation to evaluate the likelihood that the level of observed agreement between our two raters was likely to occur by chance while controlling for rater bias. As before, the probability that our data were observed by chance according to this Monte Carlo simulation was exactly zero.

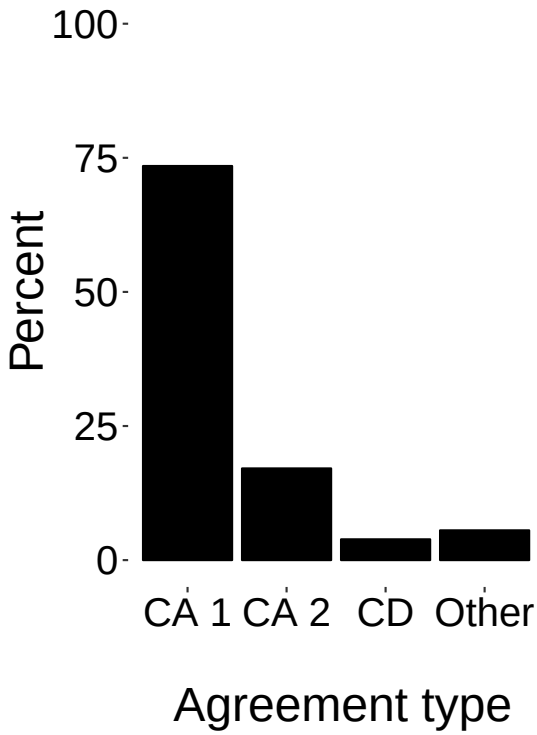

Figure 3. Distribution of ratings as a function of rater agreement in the SUBTL data. CA1, CA2: Certain Agreement that the line of dialog was linked to meaning 1 or meaning 2, respectively. CD: Raters were certain about their classifications but their classifications disagreed with one another. Other: all other cases.

As an interim summary, the broad similarity and high reliability of the FAN data and SUBTL data provides some initial evidence that the frequency of associates related to that homonym produced in a free association task, and the frequency of a particular meaning of a homonym used in natural language, have a similar distribution. Additionally, it suggests that the "local" context, be it either a single associate or a single line of subtitles, contains sufficient disambiguating information to lead to confident agreement among raters in their classification of the evoked meaning in the vast majority of cases.

Calculating estimates of meaning frequency from meaning classifications

We converted the rater classifications into estimates of the largest relative meaning frequency for each homonym, which we refer to as *biggest*, and which we use as our primary measure of meaning dominance throughout all of the subsequent analyses. In prior explicit meaning frequency rating tasks (Armstrong, Tokowicz, & Plaut, 2012; Armstrong et al., 2015), this approach has been shown to be both a simple and reliable means of probing the relative dominance of a homonym's meanings. In part, such a simple method is useful because virtually all homonyms are associated with only two meanings that are used with even a modest base frequency. Thus, measuring the largest relative meaning frequency (which typically ranges between 51% and 100%) also implicitly provides information on the frequency of the subordinate meaning, because in total these values usually sum to 100% (except for a few cases where the homonym had more than two meanings, or some classifications are not associated with any dictionary meaning). The steps in calculating *biggest* were similar but not identical for the datasets.

FAN. To compute *biggest* for the FAN data, we focused only on cases for which there was agreement on a particular dictionary definition across all raters regardless of the confidence of each classification. We first separated the associates based on the meaning classification. Next, we replaced the associates themselves with their production frequencies (e.g., the associate *money*

was produced by 115 participants for the probe *bank*). Summing these production frequencies across associates with the same meaning classification produced an estimate of the raw frequency with which each meaning was evoked in the FAN task. Finally, we selected the meaning with the largest raw frequency and computed the percentage of the total raw frequencies covered by that meaning.

SUBTL. Recall that in the SUBTL data, we extracted for rating a sample of 100 lines of subtitles for almost all homonyms, or all available lines for a few homonyms that were part of fewer than 100 lines. As for the FAN data, to compute *biggest* for the subtitles data, we included all lines which showed agreement regardless of confidence level across the two raters. We grouped these rated lines for each homonym into subgroups according to the meaning classification offered by the raters. The estimate of *biggest* for each homonym was then simply the percentage of the lines in the meaning subgroup with the most lines.

Comparison of different meaning frequency estimates

Comparison of the distribution of meaning frequency across norms. Having established that our norms have a high degree of internal validity and reliability across raters, we turn next to a comparison of our estimates of meaning dominance, as operationalized by the *biggest* measure against two other measures: the original eDom *biggest* norms derived from explicit meaning frequency estimation, and Twilley's *U*, derived from free associate classification and a slightly different operationalization of dominance.[2] The comparison between our

[2] In essence, the Twilley *U* measure reflects how confident an individual would be, on average, if they had to guess which meaning a homonym would evoke in a given context before they had actually seen the context. For homonyms with one clearly dominant meaning, that meaning would be highly likely to be evoked in any given context, so uncertainty would be low. In contrast, a homonym with two meanings of equal frequency would be associated with high uncertainty because either meaning would be equally likely to be evoked in a given context. Note that this implies a negative relationship between *U* and *biggest* estimates. For ease of comparison across different plots and datasets, we have multiplied all of the *U* values by -1 so that the relationship is positive. Note that although the two sets of computations used to derive *biggest* and *U* measures of domaince differ in the details, for words which have only two meanings, these values correlate extremely highly, $r \geq .90$, as reported in Armstrong, Tokowicz, and Plaut (2012). Thus, we interpret major differences between *biggest* and *U* as arising primarily from differences in the distribution of meaning frequencies, not to discrepancies between the mathematical formalisms underlying each

biggest estimates from the FAN data and of the Twilley et al. *U* estimates may provide particularly useful insight into the stability of measures derived from free association norms.

For our first comparison, we plotted the distribution of largest relative meaning frequencies for the intersection of the eDom, FAN, and SUBTL data. The results are presented in Figure 4. (Because the *U* dataset is so much smaller, we include the comparison of the intersection of the four datasets in our supplementary analyses in the Appendix; see Figure A3.) This figure shows a high level of similarity between the FAN and SUBTL data across the entire distribution. Like FAN and SUBTL, the eDom data also show that there are more homonyms with moderately dominant meanings (*biggest* values near 80%) than with relatively balanced meanings (*biggest* values near 50%). However, unlike FAN and SUBTL, there are very few homonyms with eDom *biggest* values of close to 100%. At first blush, these results suggest that our new FAN- and SUBTL-derived estimates are quite similar, suggesting that participants learn the frequencies of homonym meanings from exposure to natural language and then generate associates in proportion to their exposure to each meaning. However, such broad similarity at the distributional level does not necessarily reflect good agreement at the item level (see, e.g., the difference in distributional similarity vs. similarity in individual items across two dialects of Spanish; Armstrong et al., 2015). To preview some of our follow-up analyses, this initial impression does indeed appear to be an oversimplification of the data. Additionally, the presence of "homonyms" with *biggest* values at or very near 100% also suggests that identifying homonyms based on the number of entries in the dictionary may also lead to the inclusion of some words that effectively only have one meaning according to the FAN and SUBTL data. This issue is discussed in additional detail in the Appendix.

---

measure.

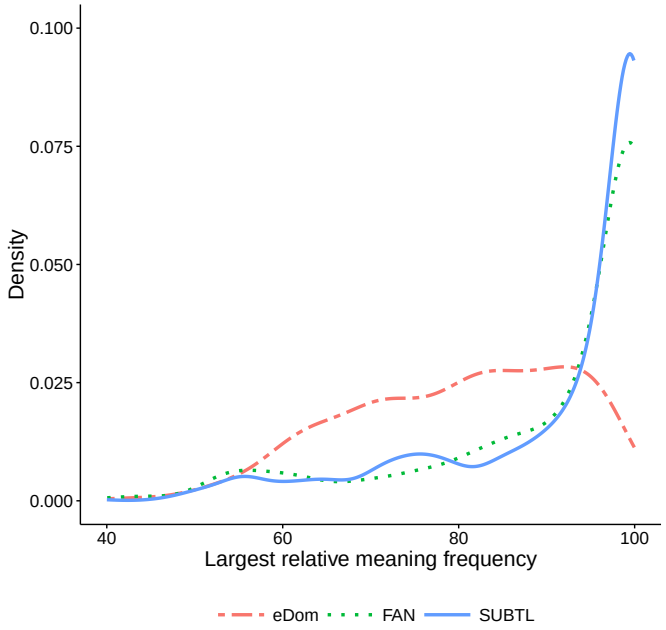

Figure 4. Density plot of the distribution of largest relative meaning frequencies for the intersection of the eDom, FAN, and SUBTL norms (n = 357). A small amount of smoothing was applied to avoid jageddness due to not having data for every x-value.

To gain finer-grained insight into the similarities and differences between the different estimates of relative meaning frequency, we next examined the scatterplots and correlations for all pairs of measures, using the intersecting items in each pair of measures; see Figure 5 (for the analogous comparisons of the intersection of the datasets, see Figure A4 in the Appendix). The figure shows that although all of the correlations were significant, their strength varied considerably, from .26 to .57. All correlations between eDom, FAN, and SUBTL were strong (between .55 and .57), which could suggest that each of these variables is tapping into the underlying "meaning frequence" construct to an equal degree, and that each measure is also tapping into a unique portion of the representation of meaning frequency. These strong relationships also highly contrast with those observed between eDom and *U*, and SUBTL and *U*, where the correlations were halved (.26 in both cases), and were very similar to the weak relationship observed between eDom and *U* in the original eDom article ($r = .27$ Armstrong,

Tokowicz, & Plaut, 2012).

Interestingly, the correlation between FAN and *U* (.45) was more similar to that between eDom and *U* and SUBTL and *U*, but nevertheless was weaker than the correlations between eDom, FAN, and SUBTL. This result is surprising given the overall similarity of the methods to produce our FAN data and those used by Twilley et al. (1994). It suggests that the *U* dataset may not be tapping into the same construct in a similar degree to the other measures. As a final observation, there was evidence in all of the plots involving our FAN and SUBTL data that those measures were generating many more datapoints at or near *biggest* values of 100, whereas the eDom *biggest* and *U* data were not strongly clustered at the upper end of the range in the scatterplots. In the discussion, we outline several possible explanations for these observations.

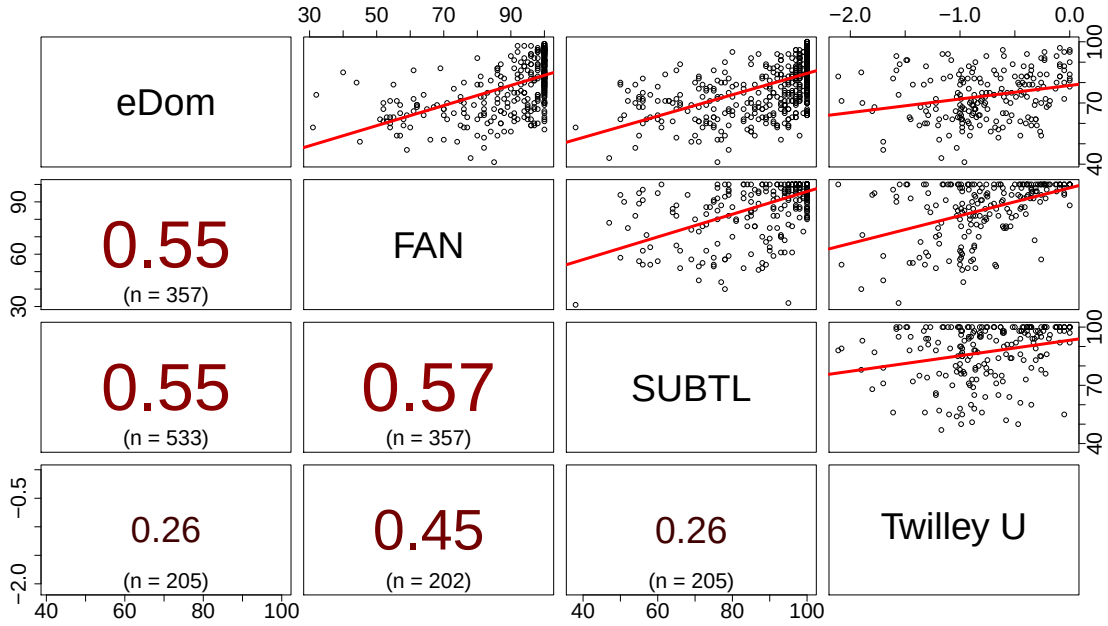

Figure 5. Combined scatterplots and correlograms (r-values) depicting the relationships between the four measures. The full set of available data for each pair of measures was included, so different numbers of observations are included in each cell. Larger and darker red (vs. black) numbers indicate larger correlations. All correlations were significant, $p < .05$, two-tailed.

Correlations with other psycholinguistic variables. The previous set of analyses indicate that despite some general agreement among the different estimates of relative meaning frequency, the correlations are still far from perfect—even the strongest correlations failed to account for more than 33% of the variance. Yet, for our new estimates and the original eDom estimates, the inter-rater reliability was always quite high. These differences must therefore reflect something systematic about how each measure taps into mental representations. One possibility is that other psycholinguistic properties of the words are influencing the rate with which different meanings are evoked. In support of this possibility, in the prior eDom studies, some weak but statistically significant relationships have been detected, typically involving variables that relate to the lexical or semantic properties of a word. Here we replicate and extend these analyses, testing the relationship between the same psycholinguistic variables and the different measures of dominance.

Figure 6 presents the results using the intersecting items in each pair of measures. (Figure A5 is the corresponding figure in the Appendix.) Several interesting observations can be drawn from Figure 6. First, all of the measures of meaning frequency are significantly correlated with other measures that relate to the representation of a word's meanings. In all cases, there is a negative relationship between meaning frequency and the number of meanings associated with a word in the Wordsmyth dictionary, indicating that words with fewer meanings tend to have a more dominant meaning relative to words with many meanings. A similar pattern, although slightly less systematic across all of our measures of meaning frequency, emerges between meaning frequency and number of senses (i.e., sub-entries in the dictionary), number of noun interpretations, and number of verb interpretations. This last case is of particular interest because of debates surrounding the separate or integrated representation of noun and verb knowledge in the mental lexicon, and we return to this point in the discussion (Mirman, Strauss, Dixon, & Magnuson, 2010). These results clearly point to the importance of controlling for these other lexico-semantic factors in studying homonymy effects, although more research will be needed to determine how

and why meaning frequencies spread out across more meanings, senses, and parts of speech.

Moving further right in the table, the similarities between the measures largely end. Only the eDom measure shows modest significant relationships between meaning frequency and most other measures, including the number of noun and adjective interpretations and imageability, a semantic property that is potentially distinct from the representation of different numbers of interpretations. The only non-significant correlations related to eDom were for variables that are sublexical in nature (number of phonemes and bigram frequency). The presence of a number of significant relationships in eDom but not in SUBTL is also particularly noteworthy because of the high number of observations in both of those cases, so those different patterns are not attributable to differences in power. The comparison of the relationship between the different psycholinguistic variables thus suggests at least modest differences between how the different meaning frequency estimates covary with estimates of other psycholinguistic properties. This, in turn, could help explain at least part of the differences observed between the different measures of meaning frequency, potentially because different tasks engage the representations of these covarying psycholinguistic properties to varying degrees. We return to this point in the discussion.

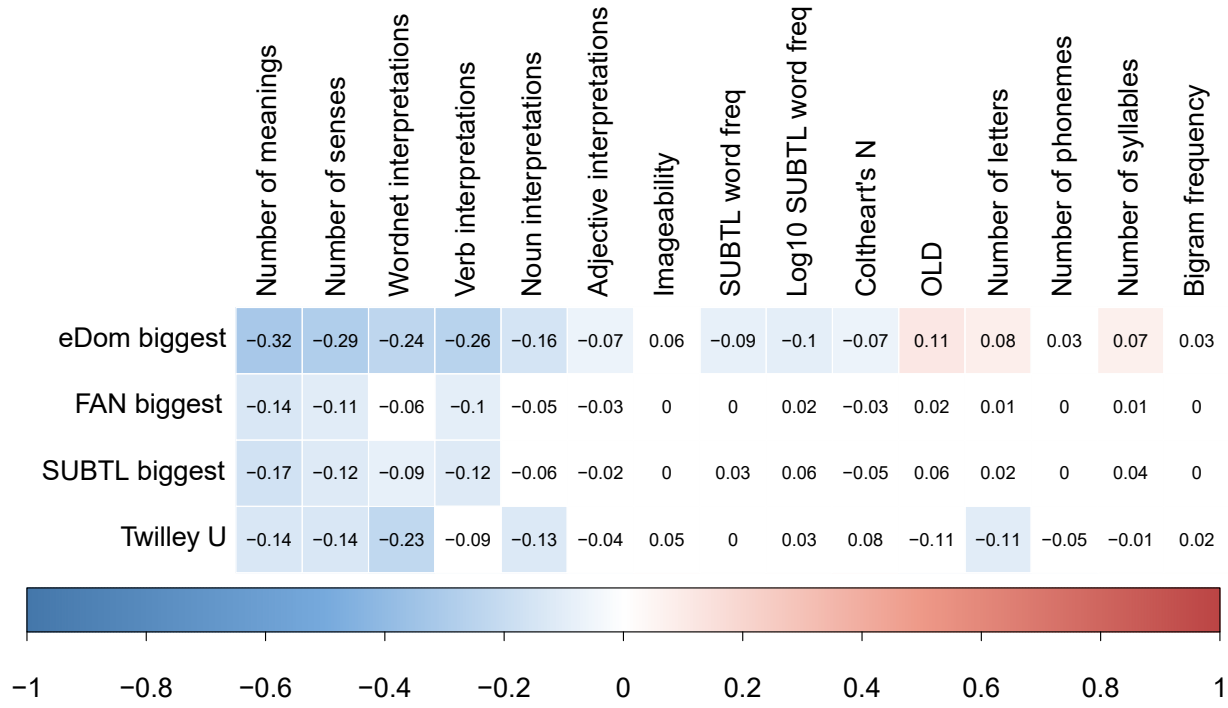

Figure 6. Pairwise correlations of the meaning frequency estimates and common psycholinguistic variables, including all available data for each measure.
($N_{eDom} = 533, N_{FAN} = 357, N_{SUBTL} = 533, N_{Twilley} = 205$). Colour shading is only included for correlations that reached significance at $p < .05$, one-tailed. Number of meanings, number of senses, and the number of verb, noun, and adjective interpretations were derived from Wordsmyth (Parks et al., 1998). The Wordnet Interpretations column denotes the total number of interpretations in Wordnet for each homonym (Fellbaum, 1998). Word frequency, and its log-transformed variant, were derived from Brysbaert and New (2009). Coltheart's N, a measure of orthographic neighbourhood size (Coltheart, Davelaar, Jonasson, & Besner, 1977), as well as the counts of number of phonemes, number of syllables, and letter bigram frequency were taken from the eDom norms (Armstrong, Tokowicz, & Plaut, 2012). Orthographic Levenshtein distance (OLD) is essentially a generalization of Coltheart's N, and counts the number of other words that can be created by letter substitution, to also include neighbours created via addition and deletion of letters; taken from Yarkoni, Balota, and Yap (2008). Note that the scales are reversed for OLD and Coltheart's N such that larger OLD scores correspond to lower Coltheart's N.

Predictive validity of measures. The previous sections established that despite the high reliability of the underlying data, the different methods for estimating relative meaning frequency yield datasets that only show moderate correlations between one another, pointing to important and systematic differences between the various measures. In the next set of analyses, we

investigated whether these systematic differences help or hurt a regression model's ability to predict performance in a few experimental tasks, which may shed some additional light on which measure(s) might align best with human mental representations of homonym meanings.

Three target datasets were identified for these analyses. The first two were the lexical decision data from the English Lexicon Project (ELP; Balota et al., 2007), which focuses on American English, and from the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012). The third dataset was the performance data from the Calgary semantic decision project (Pexman, Heard, Lloyd, & Yap, 2017), in which participants decided whether a word refers to something concrete or abstract.

The advantage of using data from the ELP and BLP mega-studies is that their broad coverage of the English language makes it likely that a large proportion of our homonyms will have been included in each study. In some respects, the ELP data are the optimal set to use in the present context — both the raw data that was normed in our studies as well as our raters themselves are all based in North America. This should minimize potential dialect differences (which have previously been shown to be non-trivial at least in some dialects of Spanish; Armstrong et al., 2015) and maximize the degree to which the same population was involved in both the derivation of our measures and the performance in an external task. However, the original eDom study found that the effects of relative meaning frequency in explaining the ELP lexical decision data were typically weak and often not significant. As Armstrong and Plaut (2016) discuss, ELP used nonwords with sublexical properties that may have biased participants to rely on lower level visual/orthographic information rather than semantic information.

To address the potential issue with ELP, we also opted to analyze data from BLP. On the one hand, these data are sampled from a different dialect of English, and therefore the variance explained by some psycholinguistic variables extracted from the American dialect of English may explain less in the British dialect of English. In particular, these psycholinguistic variables include relative meaning frequency collected across British and American dialects of English

using the eDom explicit rating method, as recently documented by Maciejewski and Klepousniotou (2016) (for additional evidence from dialects of Spanish, see Armstrong et al., 2015). On the other hand, the BLP nonwords were much more wordlike in a number of respects compared to ELP. Thus, we considered it to be an empirical question whether BLP or ELP would provide a more sensitive setting for exploring meaning frequency effects using our measures.

Using the third dataset, the Calgary semantic decision project (Pexman et al., 2017), came with both advantages and disadvantages: although this task should, in many ways, be more amenable to studying the effects of meaning frequency because it explicitly involves activating and making decisions based on the meanings of words, many fewer homonyms were available in this dataset (range: 30—91 homonyms available in a given correlation; see the table captions for details). Thus, conclusions based on this dataset must be drawn cautiously. Nevertheless, we view our explorations with this dataset as being important because they can help provide at least an initial evaluation of the viability of this type of task for the study of ambiguity effects. Moreover, convergent evidence from mega-studies other than of lexical decision are critical for making general claims regarding the mental representations of ambiguous words.

Figure 7 shows the simple correlations between the various meaning frequency estimates and the dependent measures (accuracy and reaction time) in each of the three mega-studies. (Figure A6 is the corresponding figure in the Appendix.) The most striking initial observation from the results of these comparisons is that eDom, SUBTL, and FAN all predict significant variance in ELP and BLP, whereas *U* does not; FAN and SUBTL furthermore predict significant effects in the Calgary norms. In more detail, the FAN and SUBTL measures show the highest similarity in that they were negatively correlated with RT in both BLP and ELP, such that more balanced homonyms were associated with slower RTs, and vice versa. This relationship is more consistent with typical findings in lexical decision (for discussion, see Armstrong & Plaut, 2016; Armstrong, Tokowicz, & Plaut, 2012) than the relationships observed between eDom and the lexical decision data. In a related vein, both FAN and SUBTL predict significant increases in

accuracy for homonyms with more dominant meanings in the Calgary data. In contrast, the correlations with eDom indicate that accuracy decreases in the lexical decision data. These results parallel the results obtained using simple regression reported by Armstrong, Tokowicz, and Plaut (2012). However, in that prior work the authors also found that these effects were no longer significant and were (numerically at least) typically in the opposite direction once other psycholinguistic properties were controlled for. This suggests that these simple correlations must be interpreted cautiously until the effects of other psycholinguistic covariates have been controlled for. We examine this issue in additional regression analyses reported in the following section.

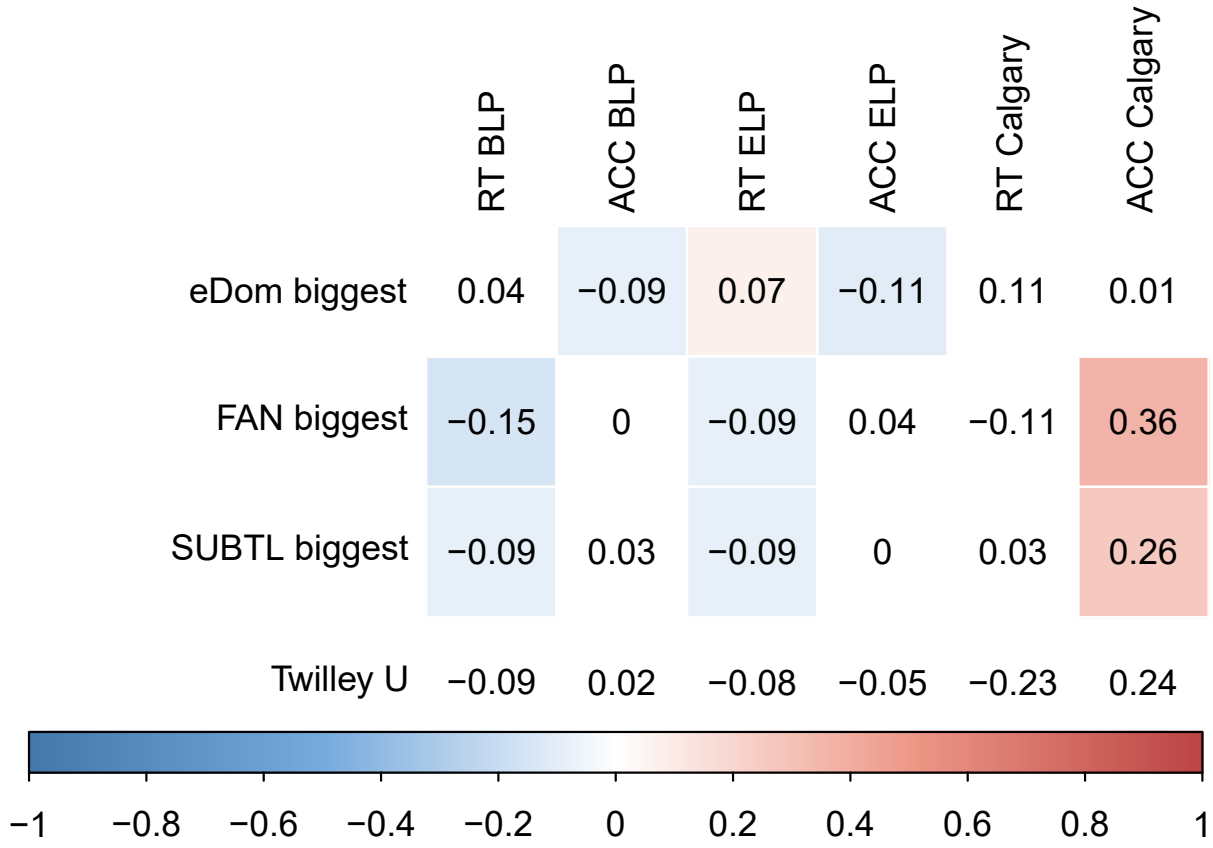

Figure 7. Correlation of meaning frequency scores and measures of human performance (accuracy [ACC] and reaction time [RT] in each of BLP, ELP, and Calgary), for all pairwise observations. Cells are shaded for significance when $p < .05$, one-tailed. The number of observations per cell were as follows: eDom and BLP = 518, eDom and ELP = 529, eDom and Calgary = 98, SUBTL and BLP = 517, SUBTL and ELP = 528, SUBTL and Calgary = 98, FAN and BLP = 350, FAN and ELP = 357, FAN and Calgary = 60, Twilley and BLP = 202, Twilley and ELP = 205, Twilley and Calgary = 31.

Next, we explored the unique variance explained by measures of relative meaning frequency after controlling for the effects of other psycholinguistic variables. To do so, we first computed the residuals from multiple regressions that used the results from each mega-study as the dependent variable (ACC or RT) and included log10 word frequency (Brysbaert & New, 2009), orthographic Levenshtein distance (OLD; Yarkoni et al., 2008), number of phonemes, number of

letters, number of syllables, number of senses, verb interpretations, noun interpretations, and letter bigram frequency as predictors. Except as cited above, all of these data were taken from the covariate data provided as part of the eDom norms (Armstrong, Tokowicz, & Plaut, 2012). We then created simple regression models that used the different measures of relative meaning frequency to predict the residuals. In essence, this corresponds to a stepwise regression wherein the meaning frequency estimate is added last. The results are presented in Table 2 for all available data for each meaning frequency estimate (Table A1 in the Appendix includes the analyses of the intersection data).

There were a number of consistent findings between the simple correlations and the regression models – namely, in both types of analyses the FAN and SUBTL measures significantly predicted both BLP RT and ELP RT. However, the remainder of the results differ somewhat from those observed in the simple correlations, primarily because fewer effects reached significance. For example, in the regression models, none of the eDom or *U* measures predicted a significant amount of variance, although there were a number of marginally significant effects. The eDom data now show several marginal effects, although the sign of these effects varies somewhat so there is no clear trend for a processing advantage or disadvantage for more balanced homonyms. Another notable difference from the simple correlations is that *U* has a number of marginally significant predictions for BLP RT, ELP RT, and Calgary RT. This pattern is surprising because *U* never predicted any significant variance in the simple correlations. No effects were significant involving the Calgary norms, which is not entirely surprising given the reduced number of observations available in that dataset.

Collectively, these results further support the notion that homonymy effects in tasks using isolated words are relatively weak and potentially subject to change as a function of the particular homonyms that enter into an analysis and how psycholinguistic covariates are controlled for. As such, it is critical that experimental work not focus only on replicating effects with small numbers of homonyms that have been used in previous studies, but rather test homonymy effects with

larger sets of randomly sampled homonyms (cf. Armstrong & Plaut, 2016). We expect that the availability of our new sets of norms, which represent a large set of homonyms selected to be suitable for psycholinguistic research, should facilitate future studies of homonymy that avoid this issue, particularly if they are coupled with methods for generating large sets of experimental items that are also matched on other potentially confounding psycholinguistic properties (cf. Armstrong, Watson, & Plaut, 2012).

Our results also suggest that, despite the task and dialectal differences in BLP and ELP, each of the four measures shows a very similar pattern of correlation across the two mega-studies. Further, the relative weakness of the correlations between the large sets of homonyms in the lexical decision tasks and the couple of strong correlations detected with the smaller set of homonyms in the Calgary norms also suggests that additional studies that emphasize the semantic aspects of processing may be needed over and above the standard lexical decision mega-studies. These types of studies may better tap orthographic and other sub-lexical and lexical aspects of word representation. For example, expanding the Calgary norms to include additional homonyms may provide a more powerful assay of the influence of semantics. More broadly, larger-scale tests of homonymy effects in tasks that have been reported to show strong homonymy effects (e.g., naturalistic reading tasks, e.g., Frazier & Rayner, 1990), although more labour intensive than isolated word studies such as lexical decision and semantic categorization and potentially tapping different aspects of meaning activation dynamics (Armstrong & Plaut, 2016), could be a further avenue for evaluating the representation of homonym meanings.

	eDom					FAN					SUBTL					Twilley				
	Estimate	SE	t-value	p-value	df	Estimate	SE	t-value	p-value	df	Estimate	SE	t-value	p-value	df	Estimate	SE	t-value	p-value	df
BLP RT	-0.20	0.13	-1.46	0.07	516	-0.42	0.12	-3.36	<.0001	348	-0.31	0.13	-2.41	<.01	515	-8.30	5.19	-1.60	0.06	200
BLP ACC	-0.0001	0.0003	-0.50	0.31	516	0.0001	0.0002	0.27	0.39	348	0.0002	0.0003	0.93	0.18	515	0.01	0.01	1.10	0.14	200
ELP RT	-0.05	0.19	-0.28	0.39	527	0.32	0.16	-1.96	0.03	355	-0.36	0.18	-2.05	0.02	526	-10.14	6.97	-1.46	0.07	203
ELP ACC	-0.0004	0.0002	-1.54	0.06	527	0.0001	0.0001	0.83	0.20	355	-0.0001	0.0002	-0.30	0.38	526	-0.0032	0.01	-0.58	0.28	203
Calgary RT	1.63	1.08	1.51	0.07	96	-0.19	0.98	-0.19	0.42	58	0.48	0.92	0.53	0.30	96	-27.59	33.03	-0.84	0.10	29
Calgary ACC	-0.05	0.11	-0.44	0.33	96	0.02	0.10	0.19	0.43	58	-0.03	0.09	-0.34	0.37	96	0.49	4.76	0.10	0.23	29

Table 2

Regression coefficients for each measure of relative meaning frequency when used to predict the residuals described in the text. Significant p-values ($p < .05$) are highlighted in green, and marginal p values ($p < .1$) are highlighted in yellow. All tests were one-tailed.

Discussion

The ability to derive "good" estimates of the relative meaning frequencies of homonyms is a major theoretical and methodological issue for advancing accounts of semantic ambiguity resolution, including explicit computational models of ambiguity resolution. In particular, good estimates would, ideally, show a high degree of inter-rater reliability, rapidly converge on stable norms without requiring excessive laboratory resources, for example, in terms of participant or rater time. They would also agree with other methods of deriving relative meaning frequency and significantly predict performance in other tasks, thereby demonstrating the external validity (generalizability) and robustness of each methodological approach. Although our results fall somewhat short of this ideal scenario, they nevertheless offer a number of important insights for theories and methods of probing the use of different meanings of a homonym. They also provide critically-lacking large-scale datasets for future empirical comparisons and computational work. We expand upon these points below.

First, our results show that each individual measure of relative meaning frequency is highly reliable when considered in isolation—both the FAN norms and SUBTL norms showed inter-rater agreement in excess of 90%, which is similar to the levels observed in the original eDom explicit meaning frequency norming study (Armstrong, Tokowicz, & Plaut, 2012; Armstrong et al., 2015). The high inter-rater reliability of these results was particularly surprising for our FAN data given the much lower inter-rater reliability reported in the related study by Twilley et al. (1994), a point which we return to momentarily. Our high reliability in the SUBTL dataset also contrasts with the lower agreement (65%) reported by Koeling, McCarthy, and Carroll (2005), who had raters label word meanings in sentences sampled from sub-sections of the Reuters corpus (Rose, Stevenson, & Whitehead, 2002). However, in that case, the authors argue that at least part of this poor disagreement is due to the high degree of polysemy in their word set, whereas we have focused on homonyms whose meanings may be more distinct and therefore easier to classify.

With such high inter-rater reliability, one might expect the various sets of norms to show a

high level of agreement. However, although there was some broad agreement among the different norms, there was also a surprising level of disagreement between the norms. This was true first in our examination of the distribution of relative meaning frequencies across the norms (see Figure 4), as well as in the scatterplots (see Figure 5). Here, even the strongest r^2 value failed to explain more than one third of the variability between measures. Perhaps even more striking, the strongest correlation was between two different methods of estimating relative meaning frequency, not between our FAN data and the norms reported by Twilley et al. (1994), despite both of those last two estimates being derived from the classification of free associates.

Making sense of why two similar measures derived from free association do not show extremely high similarity, as well as why our measure derived from free association shows much higher inter-rater reliability, has proven particularly challenging and we do not have a definitive explanation to offer. With only two datasets to contrast, it is difficult to identify which of several potential sources of variation explain these differences. One possibility is that these two sets of norms differ because of changes in how words are used to denote different meanings over time (Rodd et al., 2016; Swinney & Hakes, 1976). However, our FAN estimates were derived from the free association norms collected over a decade ago by Nelson et al. (2004), yet our raters displayed high levels of inter-rater agreement not displayed in the Twilley et al. (1994) study. Another potential source of variation is the exact set of participants that produced and rated the associates. Although we cannot make a definitive statement related to such population differences, both studies employed similar groups of university students so this source of variation seems unlikely to be responsible for the substantial differences across the datasets. A potentially more promising source of variation may be differences in the task context. For example, by focusing on norming homonyms and homographs, Twilley and colleagues' participants may have realized that most words have multiple meanings and that all meanings of the homonyms occur with a reasonable frequency. This may have been further exacerbated by the study of a particular set of homonyms that have relatively balanced meaning frequencies as compared to a broader

sample of homonyms (for details, see Figure 4 and Figure A3). Perhaps the awareness that the words have multiple meanings encouraged participants to generate classifications related to the subordinate meaning more often than in a free association task like that of Nelson et al. (2004) wherein homonyms form only a subset of all tested items. Such an explanation, although clearly speculative, would be consistent with work studying how recent experience shapes the activation of specific meanings, provided that awareness of the probe words having multiple meanings can prime all word meanings much as natural language context can prime specific word meanings (for a demonstration of meaning-specific priming, see Rodd et al., 2016). Evaluating these (and other) possibilities would necessitate additional intensive investigation, but these comparisons have at least made clear that the broader population of homonyms in natural language are more likely to have strongly dominant meanings than as suggested by the study by Twilley and colleagues and higher reliability can be achieved in some circumstances.

When comparing our new FAN and SUBTL norms to the original eDom explicit meaning frequency norms, it was also surprising to note that both new measures yielded much more left-skewed distributions and meaning frequencies near or at 100% of all cases. In line with basic views of statistical learning that describe how learners encode and internally represent the statistics of the linguistic environment (for discussion, see Frost, 2012; Frost, Armstrong, Siegelman, & Christiansen, 2015), our initial prediction was that SUBTL meaning frequencies would closely agree with the explicit meaning frequencies. Similarly, in classifying free associates, the basic assumption has been that the generation of associates should show a strong and proportional relationship with the underlying meaning frequencies. Although we did indeed find statistically significant relationships between our various estimates, at best each measure explains only about a third of the variance in one of the other measures. Clearly then, our assumptions about how these different tasks tap into the representation of meaning frequency requires some reconsideration, which could have profound implications both for theories and methods of studying meaning frequency.

Although we cannot claim to have a definitive explanation for these mismatches, in our view, three main avenues derived from theories of decision making and language representation are likely suspects and worthy of additional investigation. One avenue is concerned with how the decision system interacts with the underlying evidence to generate a particular response such as a numerical estimate of meaning frequency in a rating task. Extensive theoretical, computational, and empirical evidence has documented clear non-linearities in how representations of the actual probability distributions used as evidence to generate a response map onto the generated response (Luce, 2005; Tversky, 1967; Ungemach, Chater, & Stewart, 2009). A full review of the relevant decision making literature is beyond the present scope, but in brief the reduced number of homonyms with extremely dominant meanings in the explicit rating task of eDom versus the FAN and SUBTL norms would be consistent with an over-estimation of a low frequency event. This over-estimation may be further exacerbated by participants having some knowledge that a given homonym has two meanings (even if one of those meanings is hardly, if ever, used in natural language) and wanting their responses to reflect that knowledge. Such a strategic effect could also help explain why in a free association task only involving homonyms, Twilley et al. (1994) reported far fewer homonyms with strongly dominant meanings than we observed in our FAN norms.

A second possibility, which is not mutually exclusive with the first, is that different types of language knowledge are not represented in complete isolation to one another, so that other types of knowledge are influencing estimates, such as eDom, that are, purportedly, about meaning frequency only. For example, a classic assumption in many theories of language knowledge representation is that information about grammatical class is encoded distinctly from purely semantic information (Caramazza & Hillis, 1991). However, more recent neurocomputational models of intact and impaired language abilities have suggested that semantic information may be deeply intertwined with representations of a word's grammatical class, phonology, and other types of representations (Watson, Armstrong, & Plaut, n.d.). Interestingly, and potentially in line

with the latter claims, the explicit meaning frequency ratings in the eDom norms correlated, albeit weakly, with many more different psycholinguistic covariates—and particularly, covariates that relate primarily to semantic and/or grammatical structure, not lexical or sublexical properties—than either of our estimates derived from the classification of FAN or SUBTL data. Given the high degree of collinearity between these measures, it is not particularly straightforward to tease apart the exact contributions of these other psycholinguistic variables to each of our estimates of meaning frequency with a strong degree of statistical precision. However, the general qualitative pattern suggests that either explicit estimates of meaning frequency are spuriously influenced by these other psycholinguistic covariates, which are co-activated when completing the rating task, or these correlations are present because these different types of psycholinguistic information are not represented distinctly, but rather in a partially intertwined manner (see also similar discussion on meaning familiarity in Armstrong & Plaut, 2016).

Third, computational models of word learning have sometimes found that the fit between simulation and human performance can be improved by training items not on raw frequency, but on a log-transformed measure of frequency (e.g., Seidenberg & McClelland, 1989, though cf. Barak, Goldberg, & Stevenson, 2016). This type of improvement in data fit suggests that some scaling factor may need to be applied to raw estimates of meaning frequency, such as those from SUBTL, to better align these estimates of exposure to meaning with other estimates of the internal representations of these meanings. Our initial attempts to effect such a transformation and reconcile these data sources have yielded only limited success, however, suggesting that this avenue, while perhaps part of the overall picture, represents only a partial solution to the data mismatch issue. More broadly, this last avenue points to a more general direction for future investigation that considers the complex mechanisms that are involved in the encoding, long-term storage, and retrieval of knowledge in different contexts and experimental settings. The simplicity of our assumptions has clear advantages as a starting point, but similar assumptions in the study of related domains such as free recall suggest that our data may be shaped, to some extent at least,

by much more complex mechanisms (e.g., Tulving, 1967).

Taken together then, which of the different measures that we have discussed should, therefore, be preferred? To this question, we answer that there does not appear to be a single "best" dataset and the preferred dataset will depend upon the intended usage. If a researcher is focused first and foremost on meaning frequency per se, the SUBTL dataset would appear to have the best surface validity and parsimony with the literature on word frequency effects in recognition (e.g., Brysbaert & New, 2009). If a researcher is focused not on what meaning frequency information exists in the environment but on how meaning frequency information is internally represented and wants to avoid the complications of explicit responding and collinearity with many psycholinguistic covariates, the FAN data may be more appropriate. If a researcher is interested in the intertwined nature of meaning frequency and other psycholinguistic factors, the eDom dataset may be preferable because it may better reflect how the meaning frequency information in the environment interacts with memory systems that ensure the learning and representation of low frequency information, and how the representation of meaning frequency is intertwined with the representation of other psycholinguistic properties. If a researcher cares most about external validity and explaining variance related to meaning frequency, then the results of our tests of external validity using the BLP, ELP, and Calgary datasets might be the best basis of inferring the applicability of different measures to predicting performance in different tasks. However, even within this narrower context the results are mixed: there were no significant effects that were systematic across all tasks for any one measure, and there were trade-offs across measures in terms of showing more marginal effects (eDom, U) or fewer significant effects (FAN, SUBTL) and how well results generalized across lexical decision and semantic decision. Thus, there is no dataset that appears to be clearly superior to the others for generalizing to other tasks not studied here.

Summing up, selecting the "best" meaning frequency measure of those we have studied will require a careful consideration of how each of these measures relates to a researcher's particular

interests. Or, for the moment, a researcher who wants to hedge their bets might be best served by trying to select homonyms that have similar relative meaning frequencies across all of our norms.

Utility of norms for the development of computational models of ambiguity resolution and natural language processing

Our discussion thus far has focused on a number of insights that analyses of our new data can offer for theories and methods of semantic ambiguity resolution. However, we view those insights as value added over and above our initial goals when we started this project, when we simply aimed to amass a large amount of labeled (classified) data regarding the use of homonyms in natural language contexts (SUBTL) and in popular laboratory tasks (FAN) for use in historical and cross-linguistic/cross-dialectal comparisons of meaning ambiguity, and in computational modeling. Here, we focus on the latter application.

To date, the main computational models of semantic ambiguity in the cognitive psychology literature have focused on using artificially constructed semantic representations, often with balanced meaning frequencies for homonyms (e.g., Armstrong & Plaut, 2008, 2016; Kawamoto, Farrar, & Kello, 1994; Piercey & Joordens, 2000; Rodd et al., 2004). This has a clear advantage in offering simple, intuitive, and transparent insight into the operation of what are often already quite complex computational models. However, it also has the disadvantage of not demonstrating how these models scale to realistic representations that capture the diversity of contexts, complexity of meaning representations, and random noise encountered in behavioural data or natural language.

Before embarking on our own norming project, we also examined whether alternative sets of more realistic labeled (classified) data may have been developed in the domain of computational linguistics, which has been somewhat divorced from the cognitive psychology literature on this topic. There, we did discover data that was somewhat in line with our aims, but which we still felt was lacking in some respects if the aim is to build a tight link between human representations of word meaning and computational models. For example, several computational linguistic

resources have been developed which consist of interpretation-tagged annotations of natural text (e.g., Passonneau, Baker, Fellbaum, & Ide, 2012; Taghipour & Ng, 2015). Our main issue with using these corpora to estimate meaning frequency, however, was that they typically forfeit significant depth and amount (if any) of human annotations in order to gain breadth of coverage and total number of annotations (e.g. Taghipour & Ng, 2015). At the opposite end of the spectrum, trading forgoing breadth for depth, several corpora have been developed that provide very extensive sets of annotated data but for far fewer words. For example, corpora exist that provide $4000+$ sentences for each of the words *line*, *hard*, and *serve* annotated with the appropriate interpretation (Leacock, Towell, & Voorhees, 1993; Leacock, Miller, & Chodorow, 1998). Closer to our own work, which strikes a balance between breadth and depth, the DSO corpus (Ng & Lee, n.d.) has almost $200,000$ sentences for 191 words. That corpus, however, focused on annotating text taken from the Wall Street Journal and Brown Corpus, which may not be as representative of natural language usage as our SUBTL data derived from film and television subtitles. In addition, the DSO annotations were derived from WordNet (Fellbaum, 1998), which does not delineate between unrelated meanings and related senses in the same way that Wordsmyth does, which makes the structure of the latter particularly useful for the study of homonymy. (For additional discussion of the difficulties of sense tagging with WordNet senses, see Palmer, Babko-Malaya, & Dang, 2004.) As another example, Taghipour and Ng (2015) used WordNet senses to label the MultiUN corpus (Eisele & Chen, 2010), an assembly of United Nations Documents, which are not the primary reading material of participants who complete psycholinguistic experiments.

Relatedly, and more closely related to the present work because it also includes an examination of interpretation frequencies, several projects have developed datasets with the specific aim of studying meaning frequency distributions, but in addition to often being of a smaller scale (~50 words), these corpora also may be less representative of typical human language exposure, usage, and mental representation of meaning compared to our data. For

example, Koeling et al. (2005) developed a human-annotated corpus using up to 125 sentence usages for each of 40 words in the Reuters corpus (Rose et al., 2002). However, because one of their primary aims was to examine domain-specific distributions of meanings, they focused on strongly contrasting domains (FINANCE and SPORTS in the Reuters corpora) that cover only a portion of typical language exposure. Furthermore, they also based their classifications on Wordnet classifications (for a similar effort, see Bennett et al., 2016).

As such, although our data are certainly complementary in many respects to what has already been created in computational linguistics, we view our datasets as achieving a valuable trade-off of breadth versus depth given the current state of the literature. Our norms should also be particularly useful for facilitating the development of models that can be evaluated against large numbers of homonyms in typical language contexts (facilitated via our SUBTL data) and that can relate to experimental paradigms (facilitated via our FAN data). Our coarse measure of rater confidence and our data from multiple raters should also prove relevant for evaluating models with respect to human behavior. Already from our analyses, it is clear that models should not be considered to have "failed" in some respects if they do not achieve 100% classification accuracy. Our human raters agreed with full confidence for over 90% of the ratings. This shows a high level of reliability for our task, but also indicates that successful cognitive models would best simulate human disambiguation abilities by demonstrating similar patterns of mostly agreeing and disagreeing relative to human raters. Presumably, broader contextual information and/or sensorimotor information is used to resolve many of these ambiguities in more natural settings, which may suggest further directions for the use of models that are sensitive to word co-occurrence patterns across broad sections of text, e.g., as in LSA (Landauer & Dumais, 1997) or topic models (Griffiths, Steyvers, & Tenenbaum, 2007), and/or the development of more "embodied" semantic models (e.g., Johns & Jones, 2012). At a finer grain, our simple measure of rater confidence may also offer an additional means of assessing model classification certainty in a more continuous as opposed to binary level of analysis. For example, if a model's certainty that

a particular classification is correct covaries with a human rater's certainty on that item, this could provide an additional indication that a model represents and resolves ambiguity in a similar way to humans.

Taken together, the public availability of all of our labeled data, as well as our summary meaning frequency estimates for each homonym, all of which took several hundred rater-hours to collect before even beginning the analyses, should benefit researchers in two main ways. First, it should clearly facilitate developing better computational models from a cognitive perspective. Second, these norms should also help facilitate additional interdisciplinary interplay with other domains, such as computational linguistics, which have remained somewhat distinct from the cognitive literature on these topics and which certainly have much to offer both theoretically and methodologically. Interested readers can access all of these materials from `www.blairarmstrong.net/homonymnorms/`.

Use of the datasets. As a final methodological point, we offer some suggestions for how our datasets might be most fully leveraged for the development of models that both generate highly similar classifications (both accurate and inaccurate) to human raters, and that can generalize appropriately to a range of novel contexts. These insights derive from our examination of the computational linguistics literature, and serve to highlight the value of additional interdisciplinary research into the study of semantic ambiguity, and the fitting of human behaviour more generally.

One common approach to developing computational models of language, decision making, or other cognitive processes that can provide a quantitative fit to empirical data (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2000; Ratcliff, 1978; Usher & McClelland, 2001) is to attempt to identify an optimal set of parameters that maximize the similarity between the empirical and simulation data. This then serves as an "existence proof" that the model could explain the observed phenomena. However, such an approach contrasts with typical practice in computational linguistics, where the primary goal in parameter optimization is to maximize generalizability to

other, *unseen* data, which also tends to reduce the bias to overfit the model to a given dataset. Two approaches are used to estimate model error when applied to new data.

The first method is to denote fixed subsets of a dataset as training, validation, and test sets. For example, a typical split uses 80% of the data as training for the learning algorithm, 10% for validating feature and parameter setting selections (which can be done repeatedly), and 10% for final testing on unseen data. This approach is popular for benchmark competitions, where the participants do not have access to the test data until their models are "frozen" based on the training and validation data. However, whereas this approach in principle ensures that the trained system is tested on unseen data (providing an accurate estimate of the true error, if the test set is large enough), once results on the test data have been examined, it cannot be considered as "unseen".

The second method is to use cross-validation, splitting the full dataset of N items into k parts, and running k instances of training on $k-1$ parts and testing on the held-out kth part (where each part serves once as the held-out test set). Leave-one-out (LOO) cross-validation refers to the special case where $k = N$ – i.e., each item in the dataset serves once as the test point, when the model is trained on the other $N-1$ data items. When using cross-validation to estimate the true error of the model on unseen data, there is a bias-variance trade-off that must be considered (Hastie, Tibshirani, & Friedman, 2001; James, Witten, Hastie, & Tibshirani, 2013): Bias results when the model is trained on less data, which can lead to overestimates of the test error of a model that has access to all the data. Variance results from overlap in the training sets, such that the estimate of the test error as their mean error has high variance to the extent that the training sets are correlated. While LOO cross-validation has the advantage of having the tested models be as close as possible to the model trained on the full dataset (lowering the bias), it also has the disadvantage of maximal overlap between the training sets (raising the variance). In practice, using $k = 5$ or $k = 10$ generally provides a good balance of the bias-variance trade-off. In addition, researchers should consider stratified random sampling of the k parts, so as to balance the number of words with various properties known to impact language processing (e.g., number

of meanings, number of senses, word frequency, neighbourhood size, etc.). This method has also been shown to reduce both bias and variance (Kohavi, 1995). Stochastic optimization of stimuli offers one method of deriving such stratified samples (Armstrong, Watson, & Plaut, 2012).

Based on these considerations, for our labeled data, whenever a model does not require training on a near-exhaustive set of stimuli following LOO procedures, we suggest that researchers follow the standard practice in computational linguistics of using cross-validation to train and test models, rather than fitting parameters to the full dataset. Using cross-validation will provide the community a better sense of how well the resulting computational models would generalize to new data on words not included here, which is a necessary property of a comprehensive model of language.

Conclusion and Future Directions

Understanding how the meanings of ambiguous words are resolved is a critical component of any theory of word comprehension, and reliable and externally valid methods of estimating the relative meaning frequency of homonyms is a critical step in developing such a theory. The present work reports new relative meaning frequency estimates derived from movie subtitles and free association norms, and compares these two datasets with data from an explicit meaning frequency rating task and a previous free association study (Armstrong, Tokowicz, & Plaut, 2012; Twilley et al., 1994). All measures were highly reliable across multiple raters and showed some moderate agreement with one another, but each set of norms was also associated with unique variance. This variance potentially reflects more complex interactions between how the statistical regularities of meaning distributions are encoded, stored, and retrieved in different task settings, as well as how representations of meaning and other grammatical and semantic knowledge are stored in an intertwined fashion.

Further work teasing apart these various dimensions is clearly needed, and we expect this work and related projects studying ambiguous meaning representations over historical time and

across languages and dialects, as well as the development of computational models of ambiguity, will be facilitated by the public availability of our norms and classified data. The general success of our norming enterprise based on the use of dictionary definitions of unrelated meanings also suggests that a similar enterprise, focused on the related senses of polysemes, may also bear fruit and further help identify and characterize these finer-grained types of ambiguity. Given the even higher prevalence of polysemes in natural language relative to homonyms, and questions regarding how polyseme interpretations are represented (for a few possibilities, see Armstrong & Plaut, 2016; Col, Aptekman, Girault, & Poibeau, 2012; Hino et al., 2006, 2010; Klein & Murphy, 2001, 2002; Rodd et al., 2002, 2004; Zlatev, 2003), the present work should therefore offer a solid foundation for advancing the understanding of many aspects of semantic ambiguity.

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* 30^{th} *Annual Conference of the Cognitive Science Society* (pp. 273–278). Austin, TX: Cognitive Science Society.

Armstrong, B. C., & Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition, and Neuroscience*, *31*, 940–966. doi: 10.1080/23273798.2016.1171366

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012). eDom: Norming software and relative meaning frequencies for 544 English homonyms. *Behavior Research Methods*, *44*, 1015-1027. doi: 10.3758/s13428-012-0199-8

Armstrong, B. C., Watson, C. E., & Plaut, D. C. (2012). SOS: An algorithm and software for the stochastic optimization of stimuli. *Behavior Research Methods*, *44*, 675–705. doi: 10.3758/s13428-011-0182-9

Armstrong, B. C., Zugarramurdi, C., Cabana, A., Valle Lisboa, J., & Plaut, D. C. (2015). Relative meaning frequencies for 578 homonyms in two spanish dialects: A cross-linguistic extension of the english edom norms. *Behavior Research Methods*, 1–13. doi: 10.3758/s13428-015-0639-3

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*(3), 445–449.

Barak, L., Goldberg, A., & Stevenson, S. (2016). Comparing computational cognitive models of generalization in a language acquisition task. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2016).* Austin, TX.

Bartunov, S., Kondrashkin, D., Osokin, A., & Vetrov, D. (n.d.). Breaking sticks and ambiguities with adaptive skip-gram. In *artificial intelligence and statistics.*

Bennett, A., Baldwin, T., Lau, J. H., McCarthy, D., & Bond, F. (2016). LexSemTm: A semantic dataset based on all-words unsupervised sense distribution learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 1513–1524). Association for Computational Linguistics. doi: 10.18653/v1/P16-1143

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

Caramazza, A., & Hillis, A. E. (1991). Lexical organization of nouns and verbs in the brain. *Nature*, *349*(6312), 788.

Col, G., Aptekman, J., Girault, S., & Poibeau, T. (2012). Gestalt compositionality and instruction-based meaning construction. *Cognitive Processing*, *13*(2), 151–170.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2000). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256.

Eisele, A., & Chen, Y. (2010). Multiun: A multilingual corpus from united nation documents. In *The 7th International Conference on Language Resources and Evaluation.*

Fellbaum, C. (1998). *Wordnet*. Wiley Online Library.

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, *28*(7), 1109–1115.

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, *29*(2), 181–200. doi: 10.1016/0749-596X(90)90071-7

Frost, R. (2012). A universal approach to modeling visual word recognition and reading: Not

only possible, but also inevitable. *Behavioral and Brain Sciences*, *35*(05), 310–329. doi: 10.1017/S0140525X12000635

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality vs. modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*, 117–125.

Gale, W. A., Church, K. W., & Yarowsky, D. (1992). One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language* (pp. 233–237).

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256–281. doi: 10.1037/0096-3445.113.2.256

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc.

Hino, Y., Kusunose, Y., & Lupker, S. J. (2010). The relatedness-of-meaning effect for ambiguous words in lexical-decision tasks: when does relatedness matter? *Canadian Journal of Experimental Psychology*, *64*(3), 180–196. doi: 10.1037/a0020475

Hino, Y., Pexman, P. M., & Lupker, S. J. (2006). Ambiguity and relatedness effects in semantic tasks: Are they due to semantic coding? *Journal of Memory and Language*, *55*(2), 247–273. doi: 10.1016/j.jml.2006.04.001

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103–120.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, *32*(4), 474–516. doi:

10.1006/jmla.1993.1026

Kawamoto, A. H., Farrar, W. T., & Kello, C. T. (1994). When two meanings are better than one: Modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(6), 1233–1247.

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.

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, *45*(2), 259–282. doi: 10.1006/jmla.2001.2779

Klein, D. E., & Murphy, G. L. (2002). Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language*, *47*(4), 548–570.

Klepousniotou, E. (2002). The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language*, *81*(1-3), 205–223.

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), 1–24. doi: 10.1016/j.jneuroling.2006.02.001

Klepousniotou, E., Pike, G. B., Steinhauer, K., & Gracco, V. (2012). Not all ambiguous words are created equal: An EEG investigation of homonymy and polysemy. *Brain and Language*. doi: 10.1016/j.bandl.2012.06.007

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*(6), 1534–1543. doi: 10.1037/a0013012

Koeling, R., McCarthy, D., & Carroll, J. (2005). Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)* (pp.

419–426). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI) - Volume 2* (pp. 1137–1143). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

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.

Langone, H., Haskell, B. R., & Miller, G. A. (2014). Annotating WordNet. In *Workshop On Frontiers in Corpus Annotation* (pp. 63–69).

Lau, J. H., Cook, P., McCarthy, D., Gella, S., & Baldwin, T. (2014, June). Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL* (pp. 259–270). Baltimore, Maryland: Association for Computational Linguistics.

Leacock, C., Miller, G. A., & Chodorow, M. (1998). Using corpus statistics and wordnet relations for sense identification. *Computational Linguistics*, *24*(1), 147–165.

Leacock, C., Towell, G., & Voorhees, E. (1993). Corpus-based statistical sense resolution. In *Proceedings of the ARPA Workshop on Human Language Technology* (pp. 260–265).

Lefever, E., & Hoste, V. (2010). Semeval-2010 Task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 15–20).

Li, J., & Jurafsky, D. (2015). Do multi-sense embeddings improve natural language understanding? In L. Marquez, C. Callison-Burch, J. Su, D. Pighin, & Y. Marton (Eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1722–1732). The Association for Computational Linguistics.

Lorge, I. (1937). The english semantic count. *Teachers College Record*, *39*, 65–77.

Luce, R. D. (2005). *Individual choice behavior: A Theoretical Analysis*. Courier Corporation.

Maciejewski, G., & Klepousniotou, E. (2016). Relative meaning frequencies for 100 homonyms: British eDom norms. *Journal of Open Psychology Data*, *4*(1).

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of the effect of context in perception: Part 1. *Psychological Review*, *88*, 375–407. doi: 10.1037/0033-295X.88.5.375

Mirman, D., Strauss, T. J., Dixon, J. A., & Magnuson, J. S. (2010). Effect of representational distance between meanings on recognition of ambiguous spoken words. *Cognitive Science*, *34*(1), 161–173. doi: 10.1111/j.1551-6709.2009.01069.x

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.

Nelson, D. L., McEvoy, C. L., Walling, J. R., & Wheeler, J. W. (1980). The University of South Florida homograph norms. *Behavior Research Methods*, *12*(1), 16–37.

Ng, H. T., & Lee, H. B. (n.d.). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *proceedings of the 34th annual meeting on association for computational linguistics.*

Palmer, M., Babko-Malaya, O., & Dang, H. T. (2004). Different sense granularities for different applications. In *Proceedings of workshop on scalable natural language understanding.*

Parks, R., Ray, J., & Bland, S. (1998). *Wordsmyth English Dictionary-Thesaurus [Retrieved September 2008 from wordsmyth.net]* (Vol. 1).

Passonneau, R. J., Baker, C., Fellbaum, C., & Ide, N. (2012). The MASC word sense sentence corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC).*

Petrolito, T., & Bond, F. (2014). A survey of WordNet annotated corpora. In *Proceedings of the Global WordNet Conference, GWC-2014* (pp. 236–245).

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: Concrete/abstract decision data for 10,000 english words. *Behavior Research Methods*, *49*(2), 407–417.

Pexman, P. M., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review*, *9*(3), 542–549. doi: 10.3758/BF03196311

Piercey, C. D., & Joordens, S. (2000). Turning an advantage into a disadvantage: Ambiguity effects in lexical decision versus reading tasks. *Memory & Cognition*, *28*(4), 657–666. doi: 10.3758/BF03201255

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchinson, C., & Adler, A. (2016). The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. *Journal of Memory and Language*, *87*, 16–37.

Rodd, J. M., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245–266. doi: 10.1006/jmla.2001.2810

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, *28*(1), 89–104. doi: 10.1016/j.cogsci.2003.08.002

Rose, T., Stevenson, M., & Whitehead, M. (2002). The Reuters Corpus Volume 1-From yesterday's news to tomorrow's language resources. In *Proceedings of Language Resources and Evaluation Conference (LREC)* (Vol. 2, pp. 827–832).

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re) consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645–659. doi:

10.1016/S0022-5371(79)90355-4

Swinney, D. A., & Hakes, D. T. (1976). Effects of prior context upon lexical access during sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, *15*(6), 681–689.

Tabossi, P. (1988). Accessing lexical ambiguity in different types of sentential contexts. *Journal of Memory and Language*, *27*(3), 324–340. doi: 10.1016/0749-596X(88)90058-7

Taghipour, K., & Ng, H. T. (2015). One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning (CoNLL)* (pp. 338–344).

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *6*(2), 175–184.

Tversky, A. (1967). Utility theory and additivity analysis of risky choices. *Journal of Experimental Psychology*, *75*(1), 27.

Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition*, *22*(1), 111–126. doi: 10.3758/BF03202766

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, *20*(4), 473–479.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592. doi: 10.1037/0033-295X.108.3.550

Watson, C. E., Armstrong, B. C., & Plaut, D. C. (n.d.). Connectionist modeling of neuropsychological deficits in semantics, language and reading. *The Handbook of the Neuropsychology of Language*.

Williams, J. N. (1992). Processing polysemous words in context: Evidence for interrelated

meanings. *Journal of Psycholinguistic Research*, *21*(3), 193–218. doi: 10.1007/BF01068072

word sense disambiguation: Algorithms and applications. (n.d.).

Yarkoni, T., Balota, D. A., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971-979. doi: 10.3758/PBR.15.5.971

Zlatev, J. (2003). Polysemy or generality? Mu. In H. Cuyckens & B. E. Zawada (Eds.), *Polysemy in Cognitive Linguistics* (pp. 447–494). John Benjamins.

Appendix

Supplementary Methods and Results

Assessing Internal Validity via Monte Carlo Simulations. Given the high amount of "certain agreement" among the raters, the CA1 and CA2 bars in Figure 2 and Figure 3 can be interpreted as plotting the frequency with which the dominant and subordinate meanings of a homonym were evoked in the free association task or subtitles rating task. However, before drawing a strong conclusion in this regard we needed to rule out a simple and less interesting explanation for our present results: raters simply have a bias to indicate that the first dictionary meaning, for instance, because this is the first definition that they saw. (We did not counterbalance order of reading the definitions as in the eDom norms; for additional discussion of this potential bias, see Langone, Haskell, & Miller, 2014.) This bias to choose the first rating might artificially inflate the levels of agreement such that what appears to be high agreement may actually be due to chance. To evaluate this possibility, we conducted a Monte Carlo (MC) simulation to assess the likelihood that the observed level of agreement would be expected due to chance while controlling for rater bias. We elected to use this approach to estimating inter-rater reliability rather than a more commonly known parametric approach (e.g., Cohen's *kappa*) because our data did not meet a number of typical assumptions for this test (e.g., we had more than two raters, our data were expected to be skewed to over-represent the first dictionary definition, given that dictionary definitions are ordered according to lexicographer estimations of relative usage frequency, our data were expected to be near ceiling). Our Monte Carlo simulation is thus a non-parametric generalization of popular parametric approaches to assessing inter-rater reliability cast as a variation of a classic MC approach. Specifically, simulated datasets are constructed to match the parameters of a larger dataset, and through calculation of a test statistic for each of a large number of simulated datasets, the probability distribution of a given outcome can be estimated.

The results of our MC simulation, repeated 1000 times, are presented in Figure A1 for the free association norms data. This figure plots the observed level of agreement (red line) as well as

the chance distribution of agreement across all three raters (data = dashed blue line, fitted normal distribution = orange line), which was generated by randomly re-ordering the ratings of each rater across the homonyms (thus maintaining the overall level of bias to meaning 1 for each rater/column of ratings), and recomputing the observed level of agreement. Ratings were considered to "agree" for the purpose of this analysis if the same meaning was indicated for a given associate, ignoring whether there was high or low confidence in assigning this meaning. We also excluded the small percentage of cases in which raters "agreed" that none of the dictionary definitions were applicable because we do not know if the raters would have agreed on a specific alternative definition not listed in the dictionary. For both of these reasons, the observed agreement here is 2-3% higher than that reported in the main text, although this has no impact on our core claims. As this figure shows, the actual level of agreement was vastly higher than that of the chance distribution, and in all 10 000 simulations, none generated a level of agreement larger than that observed empirically (i.e., the p-value that we would have observed our level of agreement by chance is 0). Thus, our high levels of inter-rater agreement are clearly not attributable to a simple bias for raters to produce classifications that favour the first meaning, or due to chance alone.

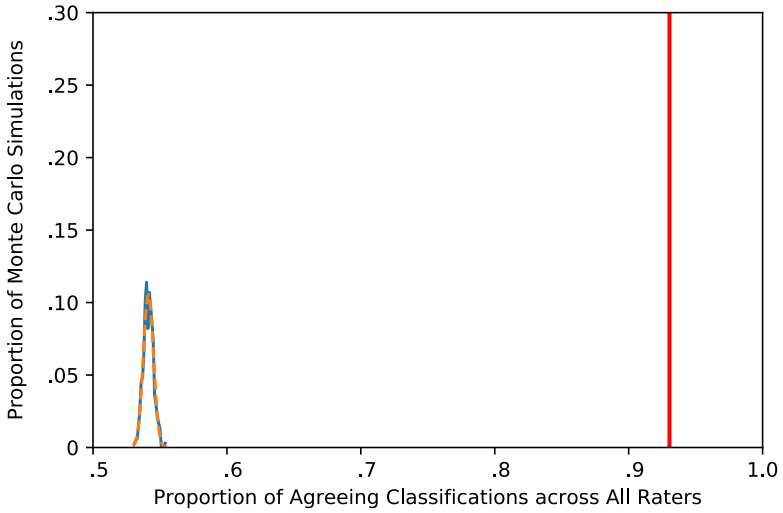

Figure A1. Distribution of agreement across all three raters in the Monte Carlo Simulations conducted on the FAN data. Red line = observed agreement. Blue dashed line = distribution of chance agreement in the simulations. Orange line = Normal distribution fit to the distribution of chance agreement.

The results of a near-identical MC simulation for the SUBTL data are presented in Figure A2. The only difference between this simulation and the previous simulation was that because there were only two ratings per line of subtitles (vsthree in the FAN data) the level of agreement produced by chance did increase somewhat. However, it still remained far below the observed level of agreement—as before, the probability that our data were observed by chance according to this Monte Carlo simulation was exactly zero.

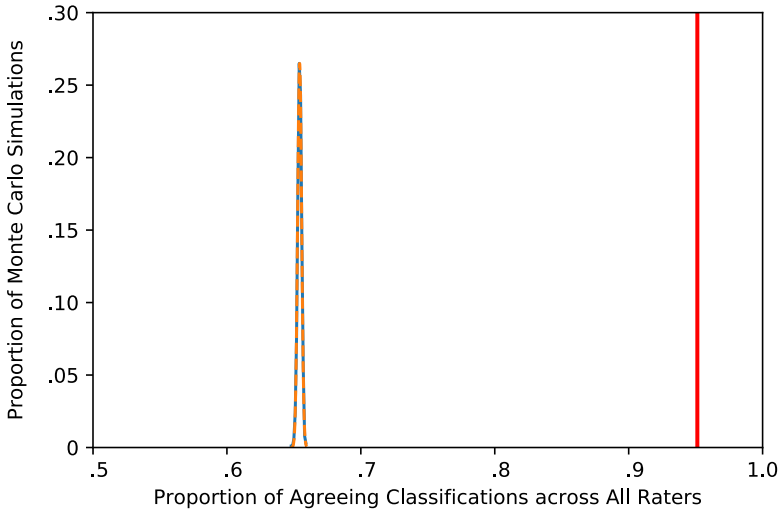

Figure A2. Distribution of agreement across the two raters in the Monte Carlo Simulations conducted on the subset of the SUBTL data having two classifications. Red line = observed agreement. Blue dashed line = distribution of chance agreement in the simulations. Orange line = Normal distribution fit to the distribution of chance agreement.

Comparison of different dominance estimates. Here we present supplementary analyses using a smaller subset of homonyms than used in the main analyses. For clarity, we call this the intersect, by which we mean only homonyms that were common to all measures included in a given analysis – a subset of 202 items. Although this set is preferable in some respects because every comparison contains exactly the same items, it should nevertheless be interpreted cautiously for two reasons. First, its size is primarily constrained by the relatively small number of items that overlap with the Twilley et al. (1994) norms. Second, as we elaborate in the subsequent sections of the appendix, the subset of homonyms in that dataset appears to not be representative of homonyms in a broader sample of the language, so the results of these comparisons may not generalize as well. (for similar findings of ambiguity effects observed only when analyzing the *U* subset, see Armstrong, Tokowicz, & Plaut, 2012).

To assist the reader in understanding where these supplementary analyses fit into the main text, we have included section headings identical to those found within the Results.

Comparison of meaning dominance across norms. We investigated whether the pattern of results reported in the main text regarding the distribution of eDom, SUBTL, and FAN data relative meaning frequencies would hold if the Twilley data were included. These data appear here in Figure A3, which presents a density plot of the distribution of the eDom, FAN, SUBTL, and U measures for the intersection of the datasets (see the corresponding Figure 4 in the main text.) Because U values are calculated on a different scale than all of our other measures, we computed a linear transform on these U values to enable comparisons in overall distributional trends (i.e., we scaled U so that it spanned the same range as the other sets and so that larger values correspond to homonyms with more dominant meanings). Two main observations are worth noting. First, in taking a subset of the data, there were far fewer cases of homonyms with meaning frequencies close to 100%; this decrease was particularly apparent in the FAN and SUBTL data, but also manifested itself to a lesser degree in the eDom data. This result suggests that the smaller subset of homonyms normed by Twilley et al. (1994) consist of homonyms with relatively more balanced meaning frequencies. This makes sense given that the homonyms used by Twilley et al. were primarily sampled from stimuli used in prior studies, and experimenters may have either an intentional and explicit bias or an implicit bias to use at least some homonyms with relatively balanced frequencies in their experiments because they may generate stronger effects of homonymy (as noted in the introduction; see also Forster, 2000, for discussion of implicit experimenter bias). In contrast, the larger set of homonyms normed in the eDom study and sampled to the largest possible extent in the new FAN and SUBTL norms comprised all homonyms in the English language that satisfied some basic length and frequency constraints relevant for psycholinguistic experimentation. In a sense then, this discrepancy suggests that the larger sets may be better at capturing the broader distributional trends in natural language and human memory representation as they relate to homonymy. This may be particularly useful with respect to developing computational models of natural language, a point which we consider in the discussion in the main text. At the same time, however, these findings, and particularly the largest

meaning frequency values of exactly 100%, suggest that some of the homonyms according to the dictionary are not actually homonyms for our participants. Thus, the use of a dictionary definitions in selecting homonyms has the advantage of being a relatively resource-efficient method for revealing how most homonyms in a language have a clearly dominant meaning, but also the disadvantage of including some items that are effectively unambiguous words at the extreme.

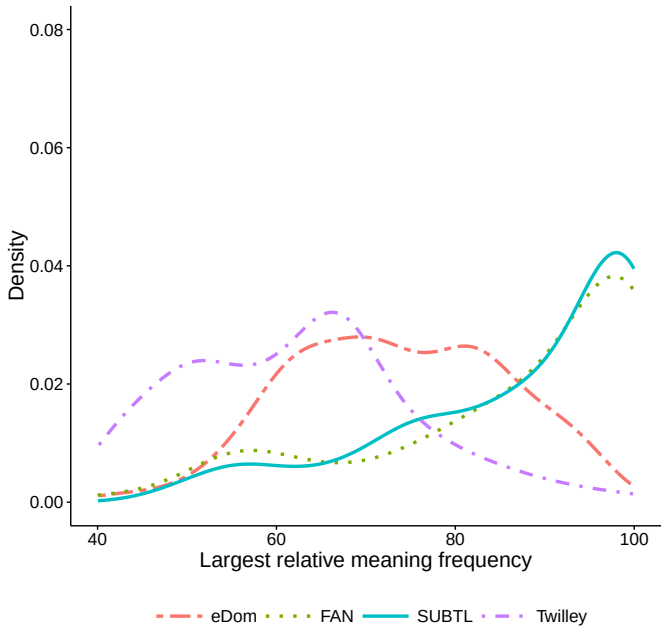

Figure A3. Density plot of the distribution of largest relative meaning frequencies for the intersection of the eDom, FAN, SUBTL, and linearly transformed Twilley U norms (n = 202). A small amount of smoothing was applied to remove jaggedness due to not having data for every point on the meaning frequency continuum.

Figure A4 describes correlations between the different measures on the intersection set of homonyms. Interestingly,the correlation values shown here vary to a modest degree (the largest change in the r-values is .10 from those shown in the corresponding analyses in the main text (see Figure 5). The most notable difference was simply a reduction in the number of extreme *biggest* values near 100. This was to be expected given our previous examination of the density plots showing that the subset of homonyms examined by Twilley et al. (1994) had fewer

homonyms with strongly dominant meanings according to all measures of meaning frequency.

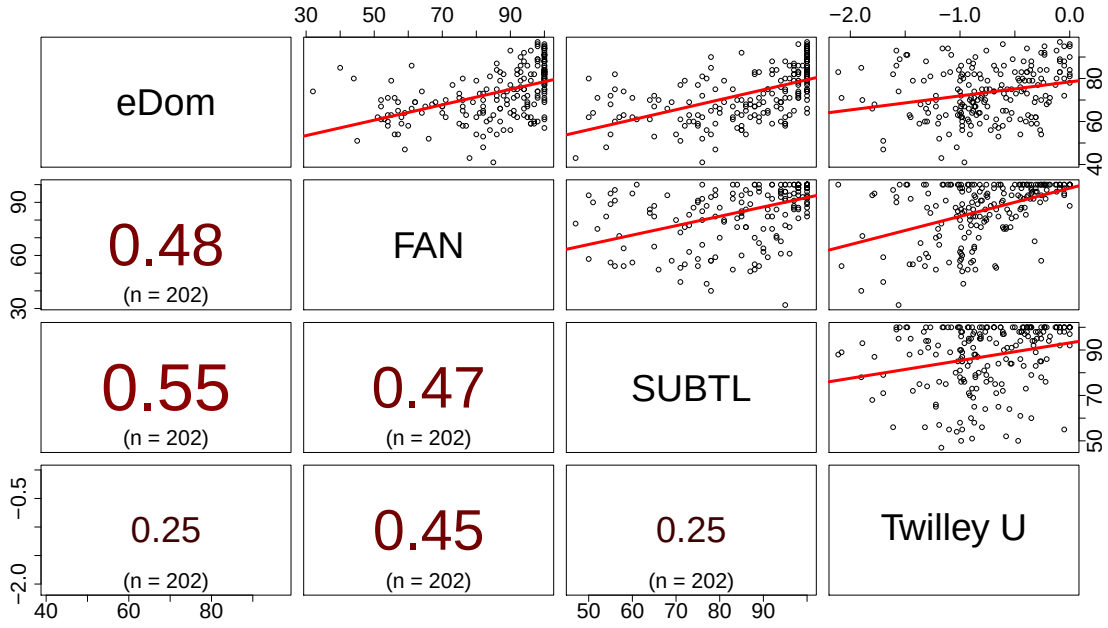

Figure A4. Combined scatterplot and correlogram (r-values) depicting the relationship between the different estimates of relative meaning frequency, as well as between relative meaning frequency and Twilley's U. Only homonyms for which meaning frequency information was available across all measures were included (i.e., it depicts the intersection of the datasets). Larger and darker red (vs. black) numbers indicate larger correlations. All correlations were significant, $p < .05$, two-tailed.

Correlations with other psycholinguistic variables. Figure A5 reports the correlations of the psycholinguistic covariates with the intersection set of homonyms. The main difference with the analyses in the main text (see Figure 6) is that a large number of previously observed effects no longer reach statistical significance, although one previously not significant effect from the main analysis does exceed the significance threshold in this case (FAN biggest and SUBTL log10 word frequency, difference in r-value between analyses = .11). This overall pattern of results is not entirely surprising given that for some of the measures the number of observations has been reduced by more than half in these analyses and most of the effects were weak even when all of the data were analyzed. However, the greater number of significant effects for *U* compared to

FAN – on the very same homonyms – further suggests that these two measures have not engaged human meaning representations in the same way despite being derived from the same kind of task, although why this would be the case is not clear.

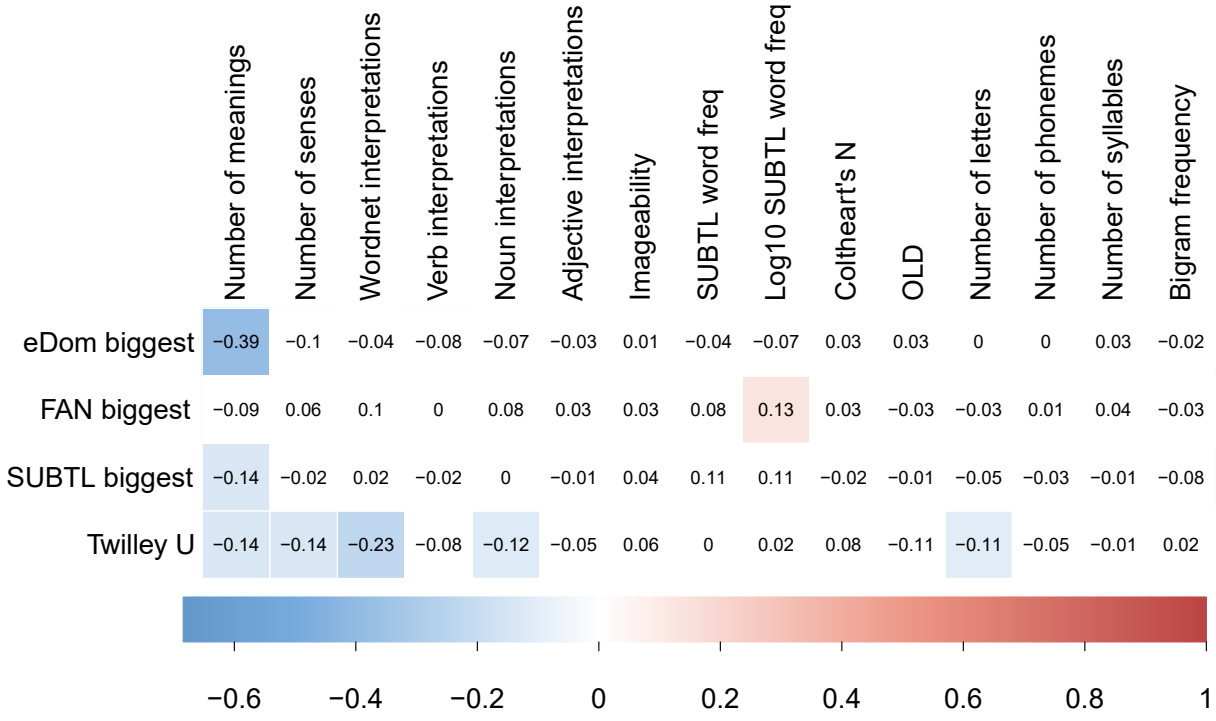

Figure A5. Correlations of dominance scores and common psycholinguistic variables, including only the homonyms that were common to all datasets (all $N = 202$). Cells are shaded for significance when $p < .05$, one-tailed. Number of meanings, number of senses, and the number of verb, noun, and adjective interpretations were derived from Wordsmyth (Parks et al., 1998). The Wordnet interpretations column denotes the total number of interpretations in Wordnet for each homonym (Fellbaum, 1998). Word frequency, and its log-transformed variant, were derived from Brysbaert and New (2009). Coltheart's N, a measure of orthographic neighbourhood size (Coltheart et al., 1977), as well as the counts of number of phonemes, number of syllables, and letter bigram frequency were taken from the eDom norms (Armstrong, Tokowicz, & Plaut, 2012). Orthographic Levenshtein distance (OLD) is essentially a generalization of Coltheart's N, and counts the number of other words that can be created by letter substitution, to also include neighbours created via addition and deletion of letters; taken from Yarkoni et al. (2008). Note that the scales are reversed for OLD and Coltheart's N such that larger OLD scores correspond to lower Coltheart's N.

Predictive validity of measures. Figure A6 presents the simple correlations between the various meaning frequency estimates and dependent measures from the three mega-studies, on only the intersecting subset of data. As discussed previously, this leads to a substantial reduction in the total number of homonyms, and as expected, reducing the number of homonyms in each analysis compared to our earlier analyses eliminated a number of significant effects (as compared to Figure 7). One new significant effect emerged, between SUBTL and ELP accuracy, and this effect followed the same counter-intuitive pattern of accuracy decreasing for more balanced homonyms observed only for eDom in the analysis of the full dataset. This difference between the full and intersecting datasets indicates that this finding is not robust, and might reflect a bias in the sample of overlapping homonyms. A number of previously significant effects noted in the previous figure simply show non-significant numerical trends in this figure, suggesting these effects remain consistent but are generally weak so as to not survive when smaller datasets are analyzed. This last point is important because it suggests that testing theoretically informative ambiguity effects like those of homonymy may require researchers to develop experiments with more observations than are often included in typical psycholinguistic studies of ambiguity to achieve a reasonable statistical power. For example, as reviewed in Armstrong and Plaut (2016), most studies of homonymy include fewer than 35 homonyms. The present analyses suggest that some effects, assuming they are real and reliable, would require substantially more observations to be detected.

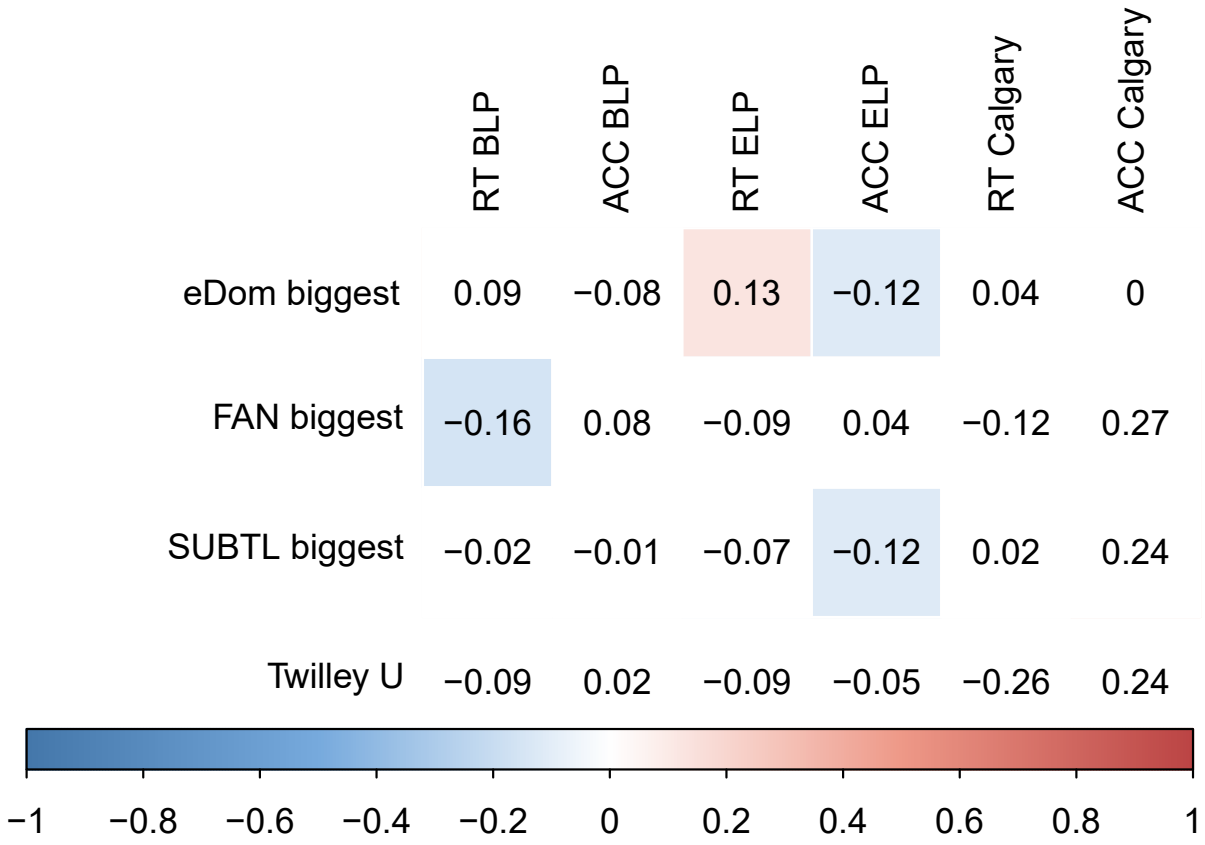

Figure A6. Correlation of dominance scores and measures of human performance (BLP, ELP, and Calgary) for the subset of homonyms present in all sets of norms. Cells are shaded for significance when $p < .05$, one-tailed. The number of observations per cell were as follows: all columns covering BLP data: 202, all columns covering ELP data: 205, all columns covering Calgary data: 31.

Next, as described in detail in the main text, we explored the unique variance explained by measures of relative meaning frequency after controlling for the effects of other psycholinguistic variables. Here, Table A1 shows the results for the intersection set, which allows us to evaluate whether any of the differences between the measures observed earlier (cf. Table 2) could have been due to differentially higher power across the measures (due to differences in number of observations) and/or effects of analyzing particular subsets of homonyms. The key difference

between the two variants of this analysis is that the effects are much weaker overall. Only one significant effect remains and except for a new marginal effect for eDom predicting ELP RT, the marginal effects either disappear or remain the same as in the main analysis.

	eDom					FAN					SUBTL					Twilley				
	Estimate	SE	t-value	p-value	df	Estimate	SE	t-value	p-value	df	Estimate	SE	t-value	p-value	df	Estimate	SE	t-value	p-value	df
BLP RT	0.16	0.20	0.81	0.21	197	-0.18	0.15	-1.23	0.11	197	0.07	0.17	0.40	0.35	197	-8.53	5.14	-1.66	0.05	197
BLP ACC	0.00	0.00	-0.69	0.24	197	0.0001	0.0002	0.59	0.28	197	0.00	0.00	-0.46	0.32	197	0.01	0.01	1.10	0.14	197
ELP RT	0.44	0.27	1.63	0.05	200	-0.06	0.20	-0.31	0.38	200	-0.09	0.23	-0.40	0.34	200	-10.43	6.99	-1.49	0.07	200
ELP ACC	-0.0003	0.0002	-1.29	0.10	200	-0.0001	0.0002	-0.42	0.34	200	-0.0005	0.0002	-2.51	<.01	200	0.00	0.01	-0.55	0.29	200
Calgary RT	1.79	1.52	1.18	0.13	28	-0.07	0.96	-0.07	0.47	28	0.67	1.16	0.57	0.29	28	-30.30	33.46	-0.91	0.19	28
Calgary ACC	-0.21	0.22	-0.94	0.18	28	0.04	0.14	0.31	0.38	28	0.02	0.17	0.11	0.46	28	0.42	4.86	0.09	0.47	28

Table A1

Regression coefficients for each measure of relative meaning frequency when used to predict the residuals described in the text. Significant p-values ($p < .05$) are highlighted in green, and marginal p values ($p < .1$) are highlighted in yellow. All tests were one-tailed. Only homonyms common to all meaning frequency estimates were entered into the analyses of each set of behavioural data.

Response Letter

Dear Dr. Yap,

Thank you for your helpful comments on our manuscript entitled "A comparison of homonym meaning frequency estimates derived from movie and television subtitles, free association, and explicit ratings", submitted for inclusion in the Society for Computers in Psychology (SCiP) special edition of *Behavior Research Methods*. We have revised the manuscript to address the comments you shared regarding our initial submission. Below, we list each of these comments from your original letter, as well as our response to each comment.

Before turning to those individual comments, we also wish to note three changes that were made to the manuscript. The first change is that we took your and the reviewers' comments regarding length to heart and have now reduced the length of the main text by 20%. This was accomplished primarily by moving some similar sets of analyses and some supplemental analyses to an appendix for the interested reader, although we did also tighten up a few sections of the text (now flagged as revisions although their core points did not change). The second minor change is that we discovered a set of up to 10 homographs that had not always been deleted from our different datasets, which we now remove. The third minor change is that in the time between the initial submission and submitting this revision we left our norming experiment for the movie subtitles running and completed a second full sweep through those ratings (as opposed to having two ratings for only 54% of the original subtitles data). These last two changes in no way alter any of our theoretical conclusions or the main patterns of the data but they do lead to many of our statistics changing very slightly (almost always between .01 and .02; only one peripheral statistic for the psycholinguistic covariates changed by more than .05). We mention these changes here simply for reasons of transparency.

In our view, these revisions have helped us make substantial improvements to the manuscript and we look forward to your comments on the revised version.

Editor Comments:

As you will see, both reviewers are positive about your paper, and I agree that this work (which should be published) is likely to yield an important methodological resource for future semantic ambiguity research. However, both reviewers provide constructive and detailed suggestions that will help make this paper even stronger and more accessible. I will not reiterate the more specific issues raised by the reviewers but I do agree that the paper is unnecessarily long now and should be made more compact (Reviewer 2). Less central analyses can be shifted to an Appendix. Reviewer 1 also makes a number of suggestions about being more nuanced in some of the claims/assertions made. I am also sympathetic to Reviewer 1's suggestion about being a little more "prescriptive" in your recommendations.

As noted above, we have made the main text more compact and moved supplemental analyses to the appendix. We provide a detailed response to the two editor comments that resonate with the specific comments of R1 and R2 below.

Reviewer 1 comments:

1. Supplementary Materials seem to be missing a sheet explaining the different variables, making the data unusable. It would also be useful to have one file with each of the norms in separate sheets.

We have adopted these suggestions and now include a variable key along with the Supplementary Materials. The data have all been combined in one file.

2. The section on inter-rater reliability is detailed and convincing, but I suggest that the authors also report a formal measure - Cohen's kappa coefficients.

We had originally considered reporting Cohen's kappa, however, we found that because we had many different raters contributing to each column of ratings, the rating agreement level was

quite high and approching the ceiling levels, and the ratings and clearly showed a bias to the first dictionary definition, we did not satisfy all of the assumptions of this parametric test. For this reason, we developed the non-parametric Monte Carlo simulation method for quantifying inter-rater reliability which is not dependent on any of those assumptions. We now explain this logic as part of the Monte Carlo section of the manuscript, which has been moved to the appendix (see page 62).

3. Page 46 - the authors argue that homonymy effects are small, such that previous studies which used a few items may have been underpowered to find such effects. The authors recommend that future studies use more items, and this is justified based on their analyses of RT data from mega studies. However, I suggest that the authors point out that this recommendation is mainly aimed at lexical decision studies. The homonymy disadvantage effect in these studies is indeed very weak and showed up only in a handful of studies that either used a highly specific set of words (Rodd's stimuli) or increased the reliance on semantics by manipulating nonword properties (Armstrong's work). Note that the effects of homonymy and relative meaning frequency are actually strong in semantically-based tasks such as semantic relatedness judgement or sentence reading. For example, Rayner's eye-tracking studies found such effects using as few as 15 items per cell. While having more items is normally a good idea, I would like the authors to recognise the limitations of their work and its implications.

We have now revised the text noted by the reviewer in two ways to address this comment. First, we note that our original comments apply to tasks studying isolated words like the lexical decision and semantic categorization mega-study data that we analyzed. Second, we acknowledge the reviewer's point that other tasks, as exemplified by some of Rayner's work, show stronger semantic effects. We avoided going too deeply into this point because of the competing pressure to reduce overall length, but the current revision should make the reader aware of the considerations that the reviewer noted (see page 38).

4. I suggest that the authors mention that their selection of homonyms has its limitations as well. While I understand that the selection, made using the Wordsmyth dictionary, was largely dictated by having collected the eDom ratings in the past, it is important to mention that there are other methods for deciding on whether a word is a homonym or not. While Rodd popularised the use of the Wordsmyth dictionary, the debate as to how to derive the homonymous status of a word is certainly not over. After all, there are people who continue to select their stimuli based on theoretical linguistics (e.g., Klepousniotou's work) or subjective ratings (e.g., Pexman's work).

We now make explicit note of the other alternative approaches to identifying and classifying word meanings listed by the reviewer on page 13 in the methods section. We also make explicit that an in-depth evaluation of which of the various methods is most appropriate was beyond the scope of the present work and that the dictionary method was chosen simply because it appears to work (regardless of whether it works "best") and is less resource intensive in terms of time or reliance on linguistic experts.

5. Likewise, the authors point out that Twilley et al.'s norms, which include a large number of balanced homonyms, are somewhat biased with respect to their stimulus selection. It would be fair to mention bias in the current study as well. Some of the words in the original eDom norms do not strike me as true homonyms. These include words whose subordinate meaning is archaic, highly infrequent, or refers to a highly specific scientific term. Having exclusively relied on dictionary entries, the eDom, SUBTL, and FAN norms include words that are homonymous to lexicographers but not to participants. In fact, most native English speakers would consider and process these words as unambiguous, and this needs to be spelled out to readers. Item ratings in Figures 6 and 7 clearly show that the eDom, SUBTL, and FAN norms include many "pseudo-homonyms" that should not be used in actual studies. While I do not think this undermines the findings of the current study, the authors should make it clear that just like with different methods of estimating meaning frequency, there are different methods of deciding on

what constitutes a homonym.

*We have revised the main text on page 27 and the appendix text on page 65 in light of the reviewer's comment. We now point out how using the number of entries in a dictionary has the advantage of identifying a broader set of homonyms and reveals more homonyms have a clearly dominant meaning than is suggested by the subset of homonyms studied by Twilley and colleagues. However, we also note that at the limit (*biggest *values of 100%, or very close to that level) this method has the disadvantage of identifying words that are homonyms according to the dictionary but not according to actual language usage as probed in SUBTL and FAN data.*

6. The authors imply that there might be differences in the estimates of meaning-frequency ratings between American and British English (page 40). I suggest that the authors read and include Maciejewski and Klepousniotou's paper that showed such differences in eDom-based ratings.

We now cite the aforementioned paper as evidence that meaning frequency estimates vary across the American and British dialects of English.

7. Finally, the authors are a bit on the fence with respect to their answer on which of the norming methods should be used. This is somewhat understandable, but readers will want a clearer statement. To me, the findings suggest that the association-based norms (FAN, Twilley et al.) seem to be the best candidate. While these two norms are not perfect and may slightly fall behind with regards to reliability, they did outperform the other norms with regards to predictive validity (esp. Twilley et al.'s ratings), and it is this property that matters the most. In word-association tasks, participants' interpretation of a probe word is biased by the dominant meaning just like it would be during on-line processing of the word. In contrast, the eDom procedure seems to overestimate the frequency of the subordinate meaning due to presentation of

the definitions (as the authors note), such that a homonym might be classed as slightly unbalanced or balanced in the norms but processed as highly unbalanced in actual word-processing tasks. The authors need to expand their discussion and make it clear that although neither method is perfect, association-based ratings might better capture bias in meaning retrieval during word processing.

We have revised our "recommendations" section of the manuscript (mainly on page 45) to try to be clearer in our recommendations. We do not repeat that entire discussion here, but in a nutshell our point is now that the "best" dataset will depend on the intended usage and theoretical assumptions of the user. If the aim is to test effects of meaning frequency per se in a way analogous to word frequency, the SUBTL data may be the best dataset. If the aim is to test how meaning frequency is encoded as an internal representation with reduced contamination from other psycholinguistic covariates, the FAN data may be best, and so on. This revised discussion should help the reader make an informed decision about which dataset is most appropriate for them to use given their aims. The revision should also help clarify our views regarding the value of association-based ratings, as noted by the reviewer and explain the circumstances under which free association norms are best but also why we decided to not make a blanket statement that they are 'best' in all circumstances.

Reviewer 2 comments:

This paper addresses an important methodological issue of how best to estimate the frequency of the meanings of ambiguous words. I found the paper interesting and informative and am certain it will be a valuable resource to researchers in this field. However, I feel the current paper is far too long and could be easily condensed. This is most clearly the case in the data sections. In addition, some of the less critical analyses/comparisons could be moved to appendices to improve readability and emphasize the most important aspects of the data.

As noted earlier in the response letter, we have taken the comments related to reducing

length to heart and have moved less critical analyses/comparisons to the appendix.

Some very minor comments:

1. Pg 10 line 21.. "we predict a higher correlation" - I wasn't sure what this correlation was being compared to. Higher than what?

We have revised the text for clarity. It now reads 'we predict a higher correlation between explicit ratings and the subtitles measure than between explicit ratings and the free association norms'

2. Pg 10 final paragraph, Experiments 3 and 4 in Rodd et al., 2016 (JML) seem relevant to this discussion of how meaning frequencies are learned.

We have revised our original text as follows to point to this research: 'According to recent empirical work, [individuals learn word frequencies] much in the same way that they learn that some words are more frequent than other words: through exposure to natural language contexts (Rodd et al., 2016).'

3. Pg 11 line 34, I think the rarity of additional definitions is not good evidence that dictionaries provide "near-exhaustive" coverage. Seems quite likely that the definitions given interfere with any lower-freq meanings making these harder to retrieve during this task.

We have revised the text in light of the reviewer's comments (see page 11). Critically, we now emphasize that is the combination of the rarity of participants offering new definitions and the fact that they do so reliably when a high frequency definition interpretation is not listed in the dictionary that is important for present purposes. Collectively, those findings from the original eDom study show that for the vast majority of words, the most frequent meanings of the word appear in the dictionary and if a high frequency interpretation is missing participants will provide

it. As such, the dictionary can be used as a reasonable basis for classification although it is not perfect.

4. Pg 21. I did not understand the Monte Carlo simulations described here. More explanation is needed. Likewise I found Figure 3 hard to understand. (Similar issues arise on page 24). Also I wasn't convinced these were central - could they be moved to an appendix?

We have now moved the Monte Carlo simulations to an appendix and revised the text reporting the methods and describing the figure for clarity (see page 62).

5. I would have found a summary table very useful (ie for each method summarise number of raters, reliability, time taken, etc.) to help compare the different approaches more directly. (More generally this approach could help greatly reduce the overall length.)

We have included a summary table on page 20 that includes several important characteristics of the tasks. We did not delve into too many detailed numerical comparisons in the table because for some tasks we do not have all of the data (e.g., Twilley and colleagues did not report the time it took their raters to rate each item) or we thought it was important to contextualize some of the numbers which was more readily accomplished in the main text.

6. Pg 25. I found the label "biggest" unhelpful. I think all you are doing is calculating the dominance of the dominant meaning. Using this more conventional terminology would be clearer in my view.

We have tried to tread cautiously on this point because it touches on a subtle but important theoretical distinction. In our view, the theoretical construct of interest is "meaning dominance" and our operationalization of meaning dominance is through the "biggest" measure. Other researchers, such as Twilley et al. also study "meaning dominance" but operationalize their

measure slightly differently in the U measure (we elaborate on this in response to a subsequent reviewer point). Additionally, the biggest *label is consistent with other related articles that we have published (Armstrong, Tokowicz, & Plaut, 2012; Armstrong et al., 2015). As such, we view* biggest *as one of several (closely related) ways of operationalizing the dominance construct and have opted to maintain the distinction, although it might slightly complicate the reading, to emphasize this distinction. To clarify this point to the readers, we have added some text to stress the different ways of operationalizing the dominance construct on page 26.*

7. Pg 27 I didn't understand why for the Twilley data the 'U' value was used rather than the equivalent dominance value that is more akin to the scores from the new data. I acknowledge that these are very highly correlated for most items (as you state the in the footnote) and that do appropriate transformations, but it still seems odd not to use the equivalent data which is available in the Twilley paper.

We were of mixed opinion on whether to report analyses using the U data or the equivalent biggest *data from the appendix in Twilley et al. As the reviewer notes, it would be more of an apples-to-apples comparison if we used the* biggest *data; however, Twilley et al. also were of the opinion that their nonlinear transform into U was warranted based on theoretical grounds (although Armstrong et al., 2012 raise some questions on this front). In the end, we opted to report the U data because we did not want to be viewed as putting free association data at a disadvantage in some of our comparisons (e.g., the tests of external validity) by choosing a different and potentially less optimal measure. As we noted in our original footnote, there turns out to be a very high correlation between U and* biggest *, ($r > .9$), and to ensure that we were not missing a major discrepancy we re-ran all of our analyses using the Twilley* biggest *measure instead of U. As expected, the same patterns were detected in both analyses. We expanded our footnote on page 26 to communicate this point in more detail, but given the pressure to reduce manuscript length we decided not to report both Twilley* biggest *and U throughout the entire*

paper.

8. Pg 51 the discussion of the idea that free association norms are not as stable as previously thought should include discussion of evidence showing that performance on this task is strongly influenced by an individual's recent experience with the words (see the Rodd paper mentioned above and several others). I'd also like to see clearer distinctions between different types of variation, eg variation over time due to changes in vocab use, within-subject variation (eg due to recent experience) and between subject variation (due to differences in individuals' idiosyncratic linguistic environments).

We have revised and added to our discussion on page 41 of some of the different sources of variation that could explain why there is less agreement with the free association norms and the Twilley and colleagues norms in particular. As the reviewer notes, there are clearly many possible sources of variation that could underlie our results. However, with only two datasets to contrast that differ in several respects, it is not possible to draw a strong inference in this regard so we aimed to strike a balance here between raising awareness of these sources of variability and the more likely sources for explaining our effects and avoiding too much speculation that would also add to the overall length.