Estimating the average need of semantic knowledge from distributional semantic models

Geoff Hollis[1]

Abstract Continuous bag of words (CBOW) and skip-gram are two recently developed models of lexical semantics (Mikolov, Chen, Corrado, & Dean, *Advances in Neural Information Processing Systems, 26,* 3111–3119, 2013). Each has been demonstrated to perform markedly better at capturing human judgments about semantic relatedness than competing models (e.g., latent semantic analysis; Landauer & Dumais, *Psychological Review*, *104*(2), 1997 211; hyperspace analogue to language; Lund & Burgess, *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208, 1996). The new models were largely developed to address practical problems of meaning representation in natural language processing. Consequently, very little attention has been paid to the psychological implications of the performance of these models. We describe the relationship between the learning algorithms employed by these models and Anderson's rational theory of memory (J. R. Anderson & Milson, *Psychological Review*, *96*(4), 703, 1989) and argue that CBOW is learning word meanings according to Anderson's concept of needs probability. We also demonstrate that CBOW can account for nearly all of the variation in lexical access measures typically attributable to word frequency and contextual diversity—two measures that are conceptually related to needs probability. These results suggest two conclusions: One, CBOW is a psychologically plausible model of lexical semantics. Two, word frequency and contextual diversity do not capture learning effects but rather memory retrieval effects.

✉ Geoff Hollis
hollis@ualberta.ca

[1] Department of Psychology, University of Alberta, P217 Biological Sciences Building, Edmonton, AB T6G 2E9, Canada

Keywords CBOW · Skip-gram · Word frequency · Contextual diversity · Likely need · average need · Needs probability · Rational analysis

Distributional semantic models (DSMs) encode the meaning of a word from its pattern of use across a corpus of text. There are a variety of classes of DSMs reported in the literature, with the two most commonly cited being latent semantic analysis (Landauer & Dumais, 1997) and the hyperspace analogue to language (Lund & Burgess, 1996), although a variety of other models have been proposed (e.g., Durda & Buchanan, 2008; Jones & Mewhort, 2007; Rohde, Gonnerman, & Plaut, 2006; Shaoul & Westbury, 2010a). Each uses different algorithms to encode the meaning of a word, but all algorithms converge on the notion that a word's meaning can be expressed in terms of a numeric vector related to its pattern of use across a corpus of text. DSMs are distinguishable from other types of vector models in terms of how they construct the vector representing a word's meaning. Other vector models of semantics represent meanings as sequences of pseudorandom values (e.g., Hintzman, 1988) or as sequences of semantic features derived from human judgments (e.g., McRae, Cree, Seidenberg, & McNorgan, 2005). In contrast to other classes of vector models, DSMs provide formal descriptions of how learning can be used to derive knowledge from experience; a word's vector representation is the output of such learning.

The field of natural language processing (NLP) has a research area called word embedding that runs parallel to work on DSMs in psychology. Word embeddings are feature vectors that represent a word's meaning, learned by predicting a word's use over a corpus of text (e.g., Mikolov, Chen, Corrado, & Dean, 2013). The main algorithmic cleave that separates NLP and psychological models is that that psychological models primarily arrive at a representation of word

Published online: 13 July 2017

meaning by counting contexts of occurrence and NLP models primarily arrive at a representation by predicting contexts of occurrence (for a comparison of these models in terms of *predict* and *count* descriptions, see Baroni, Dinu, & Kruszewski, 2014).

Two promising models from the NLP literature are the continuous bag of words (CBOW) model and the skip-gram model (both described in Mikolov, Chen, et al., 2013). Both use neural networks to learn feature vectors that represent a word's meaning. The models accomplish this by finding statistical regularities of word co-occurrences that predict the identity of missing words in a stream of text. In the case of CBOW, a contiguous string of words is presented to the model (e.g., a sentence) with one word missing. The model then has to use the other words available in the sentence to predict the missing word. The skip-gram model solves the converse problem: given a single word as a cue, which other words are likely to be present in the context?

Predict models like CBOW and skip-gram tend to better capture variation in human behavior than the count models that psychology more commonly uses. Specifically, skip-gram and CBOW have unprecedented accuracies on analogical reasoning tasks (e.g., Mikolov, Chen, et al., 2013) and better fit lexical access times and association norms than count models (e.g., Mandera, Keuleers, & Brysbaert, 2017). More broadly, predict models have superior performance on a wide range of tasks, including categorization, typicality judgment, synonym judgment, and relatedness judgment among, others (Baroni et al., 2014). CBOW and skip-gram were never developed for the purpose of providing insight to psychological problems, yet they fit the aforementioned behavioral data better than the established models in the field. There is some literature describing the relationship between predict and count models in terms of the computational problems they are solving (e.g., LSA and skip-gram are computationally, formally equivalent for certain parameter settings; Levy & Goldberg, 2014), but no literature attempting to provide an explanatory theory as to why, as a general rule, CBOW and skip-gram models perform so much better than traditional psychological models at fitting human behavioral data, other than the fact that they can be trained with larger corpora (however, see Mandera et al., 2017, who still find differences in performance when corpus size is held constant). Attempting to understand why CBOW and skip-gram models fit behavioral data so well could help inform theories of language learning and word meaning.

A possible answer comes from Anderson's rational analysis of memory (J. R. Anderson, 1991; J. R. Anderson & Milson, 1989; J. R. Anderson & Schooler, 1991). Like other aspects of human behavior and physiology, memory has been shaped over deep evolutionary time by pressures of natural selection. It is a reasonable assumption that memory retrieval should display evidence of being optimized with respect to facilitating effective interaction within the environment (see also Norris, 2006, for discussion on the optimality assumption of cognitive processes). Thus, Anderson starts from the question: What would the behavior of a memory system designed to facilitate effective interaction with the environment based on perceptual input look like?

A core prediction of Anderson's analysis is that access to information in memory should be prioritized based on its likely relevance, given an agent's current context (J. R. Anderson & Milson, 1989). This idea can be formalized with Bayesian statistics: Given a history of experience and a current set of contextual cues, the particular piece of knowledge that is most likely to be needed should be most readily accessible. Anderson and Milson refer to a memory's relevance, given contextual cues, as its needs probability (J. R. Anderson & Milson, 1989, Equation 5; see also Steyvers & Griffiths, 2008, for an accessible introduction to needs probability). Others have invoked a related construct, likely need (Adelman, Brown, & Quesada, 2006; Jones, Johns, & Recchia, 2012), which is the needs probability of a word averaged across a distribution of possible contexts.

There are clear relationships between the optimization criterion of memory retrieval, as described by J. R. Anderson and Milson (1989), and the learning objectives of CBOW and skip-gram algorithms. The learning objective of CBOW in particular has a direct correspondence to Anderson's concept of needs probability: CBOW learns to predict the identity of an omitted word from a phrasal context, given the identities of nonomitted words. Skip-gram learns to solve the converse problem: Given a word, what other words are likely to be present in the omitted phrasal context? These casual descriptions of learning objectives are supported by their proper formal descriptions in terms of conditional probabilities (e.g., Bojanowski, Grave, Joulin, & Mikolov, 2016). It is possible that CBOW and skip-gram are solving the same, or a similar, computational problem as human memory as described by Anderson and Milson. If this were the case, CBOW and skip-gram should produce behavior that is consistent with predictions made by the rational analysis of memory.

Testing predictions of needs probability with lexical access times

Estimates of needs probability should be linearly related to behavioral measures of memory retrieval times (J. R. Anderson & Milson, 1989). Within the domain of psycholinguistic research, this includes measures such as word naming time and lexical decision time, of which there are multiple large data sets freely available for simulating experiments (e.g., Balota et al., 2007; Keuleers, Lacey, Rastle, & Brysbaert, 2012). However, in such tasks, context cues are not provided, with the exception of special cases (e.g., priming experiments). For data from naming and lexical decision data sets to be usable for testing predictions of needs probability, some additional steps need to be taken.

It is important to recognize that even if contextual cues are not explicitly presented during memory retrieval tasks, such as word naming and lexical decision, participants' minds are not empty vessels. Moment by moment, thoughts are likely acting as context cues for memory retrieval in otherwise context-sparse tasks favored by psychological research. So, how do those thoughts distribute? With a reasonable answer to that question, the expected need of a word could be estimated by averaging its needs probability over the distribution of context cues supplied by participant thoughts. We use the term *average need* to mean the needs probability of a word averaged over a distribution of possible contexts and distinguish it from *needs probability*, which is the probability that knowledge will be needed, given a set of actual context cues. The abovementioned definition of average need is consistent with what Adelman et al. (2006) and Jones et al. (2012) mean by likely need: the prior probability that a word will be needed when specific contextual details are unspecified. Although it is conceptually identical to likely need, we instead adopt the term average need for two reasons. First, although Anderson has consistently been cited as the source of the concept of likely need, there is not a single use of this term in any of his relevant work (J. R. Anderson, 1991; J. R. Anderson & Milson, 1989; J. R. Anderson & Schooler, 1991). Breaking from use of likely need is an attempt to correct a misattribution. Second, we specify the exact formal relationship between an actual construct introduced by Anderson et al., *needs probability*, and our own term, *average need*.

There are two obvious and simple approaches to estimating the distribution of context cues in lexical access tasks. One is to model the distribution based on words presented on previous trials. This is based on the assumption that when a participant reads a word, he or she involuntarily anticipate other words that may follow. For an experiment that uses random presentation order, that distribution is the same as the distribution of words about which participants make decisions. The second approach is to estimate the context distribution from the topics that people talk about in their day-to-day lives. For instance, the linguistic contexts captured by text corpora extracted from Internet discussion groups (e.g., Shaoul & Westbury, 2013), entertainment media (e.g., Brysbaert & New, 2009), or educational sources (e.g., Landauer, Foltz, & Laham, 1998). There exists other work on measuring needs probability within psycholinguistic research (e.g., semantic distinctiveness count, Jones et al., 2012; contextual diversity, Adelman et al., 2006). Past work favors corpus-based approaches for estimating context distribution.

Of note, corpora that closely match day-to-day linguistic exposure typically produce the highest quality estimates of lexical variables (e.g., word frequency), as measured by the ability of those variables to predict lexical access times (e.g., Brysbaert et al., 2012; Keuleers, Brysbaert, & New, 2010; Herdağdelen & Marelli, 2016). This observation is convergent with the earlier claim that the context distribution for memory retrieval during context-sparse psychological experiments is likely supplied by the active thoughts and knowledge that participants bring with them into the laboratory; in a moment, an argument will be provided that word frequency and contextual diversity are fundamentally related to each other as operationalizations of average need.

Measuring average need

Calculation of the average need of a word involves taking an average of the needs function over the context distribution of memory retrieval. Consider the calculation of contextual diversity (CD), which is the number of documents in a corpus within which a word appears. A particular document within a corpus can be thought of as a possible context. The full corpus provides a distributional model of contexts. The calculation of needs probability, given a context, is a binary check of whether the word is present or absent in that particular context. CD scores are the sum of a binary needs function over the distribution of possible contexts. In this case, a sum would be proportionally related to an average.

Longer documents like novels and plays undergo plot developments. These developments continually necessitate the need for new knowledge as the document progresses. Empirically, this means that word frequencies are not stable across sequential slices of a document (Baayen, 2001). Since the content of discussion can change across the length of a document, a question is merited: Is context better defined in terms of a textual unit smaller than "document"[1]? Perhaps better CD measures could be produced if contexts were defined in terms of chapters, sections, paragraphs, sentences, phrases, or individual words. In the case where context is defined in terms of the smallest of these units, an individual word, CD measures would be identical to word frequency (WF) measures. Or stated alternately, if CD is a measure of average need, so is WF. What distinguishes the two measures is how they choose to define context when calculating needs probability.

CD and WF can both be thought of as measures of average need, each applying a different definition of context. In light

[1] Landauer and Dumais (1997) popularize the use of documents as units of context for DSMs. They explicitly note that for the corpus they employed, individual documents were rather short—one or two paragraph samples from K–12 textbooks. Given the brevity and specialized content of these "documents," it was a reasonable assumption that each document spanned no more than a single coherent and continuous semantic context. Explicit in this example is the idea that finding the proper unit of measurement for a context matters; larger units of measurement may pose a problem for properly measuring the semantic features present in a context.

of this, the finding of Adelman et al. (2006) that CD provides a better fit to lexical access times than WF can be can be restated as follows: Defining context in terms of larger linguistic units (documents), rather than smaller linguistic units (words), provides estimates of needs probabilities that more strongly correlate with lexical access times. However, it remains an open question as to whether intermediary definitions of context like sentences or paragraphs provide a better fit than the two extreme context sizes. The answer might be yes if documents are so coarse that they encompass shifts in the topic of discussion.

The concept of CD has recently been extended by Johns, Jones, and colleagues (Johns, Dye, & Jones, 2016; Johns, Gruenenfelder, Pisoni, & Jones, 2012; Jones, Johns, & Recchia, 2012). Their observation is that defining CD in terms of document counts overlooks the fact that some documents are not referencing unique contexts, either due to document duplication or due to multiple documents all referencing the same event (e.g., as may happen across news articles all discussing the same world event). Jones et al. (2012) propose the concept of semantic distinctiveness: the degree of variability of the semantic contexts in which a word appears. One of the ways semantic distinctiveness has been operationalized is with what Jones et al. (2012) call semantic distinctiveness count (SD_Count). SD_Count is a document count metric that is weighted by how semantically unique the document is. SD_Count is a better predictor of lexical access times than both CD and WF (Jones et al., 2012), and in an experiment using an artificial language, participant recognition times of nonce words are sensitive to effects of SD_Count when WF and CD are controlled for. Jones and colleagues have also pursued a variety of model-based approaches to operationalizing semantic distinctiveness, with similar results (e.g., Johns et al., 2016; Jones et al., 2012).

Jones et al.'s (2012) SD_Count is a calculation of average need that uses a binary needs function, defines context in terms of documents, but also includes a term that downweights the contribution of common contexts when calculating average need over the context distribution.

Measuring needs probability from word embeddings

One of the key features of word embeddings produced by CBOW, skip-gram, and other DSMs is that they allow for the measurement of similarity of meaning between two words (typically with cosine similarity between embeddings). J. R. Anderson (1991) has argued that similarity between context cues and memory features will be proportional to needs probability (see also, Johns & Jones, 2015; Hintzman, 1986). Thus, DSMs can also offer an operationalization of the average need of a word: the similarity between a word vector and a context vector, averaged over all contexts supplied by a context distribution.

Theoretical summary and research objective

Anderson's rational analysis argues that memory is always engaged in a process of forward prediction. Contextual cues that are available now are used to predict and make accessible knowledge that will be needed in the near future. The purpose of such forward prediction is to aid efficient and adaptive behavior within the world (see also Clark, 2013, for a related discussion on the biological plausibility and adaptive significance of forward prediction).

A core theoretical construct that arises from Anderson's rational analysis of memory is needs probability—the probability that some piece of knowledge will be needed in the future, given a set of contextual cues that are available now. Anderson expects that traces in memory should be accessible according to their needs probability. Thus, we arrive at a criterion for testing memory models: needs probabilities derived from them should be linearly related to empirical retrieval times. In cases where the context for forward prediction is unknown, we can expect that retrieval times should be linearly related to the average needs probability over the distribution of possible contexts.

There is currently a lack of understanding as to why CBOW and skip-gram models better fit behavioral data, including retrieval times (Mandera et al., 2017), than the DSMs more commonly used within psychological research. We hypothesize that CBOW and skip-gram are solving the same, or a similar, learning objective as human memory (as specified by Anderson's rational analysis).

CBOW and skip-gram both learn to represent word meanings as connection weights to (skip-gram) or from (CBOW) the hidden layer of a three-layer neural network. That neural network is used to predict the words most likely to appear, given information about a surrounding context (CBOW), or the most likely contexts of occurrence, given a word (skip-gram). The functional description of CBOW appears most consistent with Anderson's rational analysis of memory, whereas skip-gram appears to be solving a converse problem. As a side note, because of the formal equivalency between skip-gram and LSA (Levy & Goldberg, 2014), this also poses theoretical problems for LSA within the context of Anderson's rational analysis.

The motivation of this research is to understand why CBOW and skip-gram do so well at fitting behavioral data, particularly in lexical retrieval tasks. We believe Anderson's rational analysis of memory provides an appropriate theoretical framework for beginning to interpret the behavior of these models. Thus, we arrive at the main prediction of this research: insofar as they are plausible models of lexical semantic memory, measures of average need derived from skip-gram and CBOW should be linearly related to lexical retrieval times. Furthermore, average needs derived from these models should also be strongly correlated with other operationalizations: log WF, log CD, and log SD_Count.

Models that bear on the process of lexical access are typically concerned with the way in which stimulus properties activate lexical entries or their meanings (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; McClelland & Rumelhart, 1981; Seidenberg & McClelland, 1989). Neither skip-gram nor CBOW have any information about stimulus properties, and they do not address issues of lexical access at this level. Rather, they learn from the contingencies between words and contexts (which is determined by other words). Thus, they only bear on the question of how context determines the activity of items in memory. This activity is not the lexical access of a word's meaning, given the orthographic or phonological details of that word. Rather, it is the posterior accessibility of a lexical entry, given information about context. Ultimately, an incomplete but important piece of the puzzle of lexical access.

To begin to interpret a model within the framework of Anderson's rational analysis of memory, an operationalization of needs probability is required from that model. Needs probability is conceptually a matter of trace activation, given a probe. We thus borrow from traditions in vector models of memory and define trace activity, given a probe, in terms of vector similarity between trace and probe (e.g., Hintzman, 1986; Johns & Jones, 2015). Although they do not provide an actual model implementation of their theory, J. R. Anderson and Milson (1989) also suggest this exact approach to operationalizing needs probability.

Experiment 1

In this experiment, we derive measures of average need from CBOW and skip-gram word embeddings. We use various text corpora as distributional models of context. Results of Adelman et al. (2006) suggest that when calculating needs probability, it may be more appropriate to use documents as the unit of context rather than words. However, the gains are marginal. As discussed earlier, WF and CD are calculations of average need that employ the same needs function, but different definitions of context (WF uses a word definition, CD uses a document definition). Within the corpora employed in this study (introduced shortly), WF and CD are nearly identical (all *r*s > .99). For the current study, we define context in terms of individual words. We do this because it results in a more straightforward calculation of needs probability.

The needs probability of a memory trace can be estimated by the similarity between its structure and the structure of the probe supplied by contextual cues (J. R. Anderson & Milson, 1989). In terms relevant to DSMs, the needs function for a target word (memory trace), given a context word (probe), is the cosine similarity between the two words' embeddings. Similar equivalencies are also made in vector-based exemplar models of memory (e.g., Hintzman, 1986; Johns & Jones, 2015, where vector similarity is used to determine the activation level of a memory trace, given a probe). In the current application, the cosine similarity between a probe word (which is a stand-in for the semantic details of a possible linguistic context) and a lexical entry specifies the needs probability of that lexical entry in that particular linguistic context.

With a needs function and a distributional model of context defined, the average need of a word can be calculated as the average similarity between a target word's embedding and the embeddings for all other words in the context distribution, weighted by the frequency of occurrence of the context word. Frequency weighting is required to ensure that measures of average need are based on the proper distributional form of contexts.

Words with high cosine similarity to other words on average should be precisely the words that are needed across a broad range of contexts. Conversely, words with low cosine similarity to other words on average should be those words that are used in a very narrow range of uncommon contexts. Using as example one of the models tested in the current experiment (CBOW, trained on a Subtitles corpus), high average need words included terms like *he's, won't, that's, tell, thanks,* and *gonna*. Low average need words included terms like *cordage, shampooer, spume, epicure, welched,* and *queueing*. It is far more often, and across a more varied types of contexts, that we need to be prepared to know about telling, thanking, and doing than welching, queueing, and shampooing.

We make three predictions. First, estimates of average need should be linearly related to measures of word naming time and lexical decision time. This is the primary prediction of the current experiment and is supplied by J. R. Anderson (1991). We note that this is a prediction that other operationalizations of average need consistently fail to satisfy; CD (Adelman et al., 2006) and SD_Count (Jones et al., 2012) are logarithmically related to measures of lexical access.

Second, CBOW and skip-gram average needs should be related to other operationalizations insofar as they are psychologically plausible models of semantic knowledge. We focus on comparing values to measures of CD (Adelman et al., 2006) and WF. We exclude SD_Count (Jones et al., 2012) from analyses. SD_Count is a computationally costly calculation and does not scale to large corpora. Model-based approaches to calculating semantic distinctiveness are available that are claimed to scale to larger corpora (e.g., Johns et al., 2016); however, we do not yet have a working implementation of one such model and thus sidestep comparisons to model-based estimates of semantic distinctiveness from this analysis. Since CD and WF measures are best fit to lexical access times with a logarithmic function, it is predicted that average needs derived from CBOW and skip-gram will be linearly related to the logarithm of these two measures.

Third, the ability for average need to fit lexical access times or CD/WF measures will depend on the quality of the word embeddings that average needs are calculated from. With a poor model of semantic knowledge, we should expect correspondingly poor estimates of when knowledge is needed and when it is not. We anticipate this effect being expressed in two ways: one, measures of average need should have no predictive validity when randomized word embeddings are used. Two, as corpus size increases, average needs should improve in quality; as corpus size increases, more differentiation in word usage exists. Consequently, word embeddings likewise become more differentiated in values.

Method

Corpora

We repeat our analyses on three corpora to ensure generalizability of results. We use a movie subtitles corpus (Subtitles) that has previously been used to compare the performance of CBOW and skip-gram on modeling human performance on psychological tasks (Mandera et al., 2017), a 2009 dump of Wikipedia (Wikipedia) that is widely used in natural language processing research (Shaoul & Westbury, 2010b), and the Touchstone Applied Science Associates corpus (TASA; Landauer, Foltz, & Laham, 1998). These corpora were chosen because they vary in size (TASA 10 M tokens, Subtitles 356 M tokens, Wikipedia 909 M tokens), and all have a history of previous use in psycholinguistic research and distributional semantic modeling. The movie subtitles corpus, specifically, was chosen because WF measures derived from movie subtitles account for more variation in lexical access times than WF measures from other common corpora (Brysbaert & New, 2009). The context distribution being supplied by the subtitles corpus is likely a good match for the actual context distribution of memory retrieval during lexical access tasks. We further created two subsets of the subtitles corpus—one containing half of the full set of documents (approximately 180 M tokens; Subtitles-180 M) and another containing a quarter of the full set of documents (approximately 90 M tokens; Subtitles-90 M). Subsets were generated by randomly sampling documents from the full corpus (without replacement) until the desired token count was reached. Subtitles-90 M is a nested subsample of Subtitles-180 M.

Prior to analysis, each corpus was preprocessed by removing all punctuation and converting all words to lowercase. Words were not lemmatized (i.e., *kick*, *kicking*, and *kicks* were all treated as unique types) on the basis that nonlemmatized frequencies better account for variation in lexical access than lemmatized frequencies (Brysbaert & New, 2009).

Calculations of average need

Measures of average need were calculated from CBOW and skip-gram models after unique instances of each model had been trained on each of the five corpora. Thus, we derive average needs from 2 (model type) × 5 (corpus) = 10 models. The context distribution used for calculating average need was limited to words with entries in the English Lexicon Project (ELP; Balota et al., 2007), the British Lexicon Project (BLP; Keuleers et al., 2012), the Warriner et al. affective norms set (Warriner, Kuperman, & Brysbaert, 2013), the Brysbaert et al. concreteness norms set (Brysbaert, Warriner, & Kuperman, 2014), or the Kuperman et al. age of acquisition norms set (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). The scope of analysis needed to be limited to keep it tractable. These data sets contain entries for most words that would appear within psychological research contexts, so we deemed it a good inclusion criterion. The union of these datasets contains 70,167 unique words that will act as contexts for calculating average need.

For each corpus, CBOW and skip-gram word embeddings were trained for words appearing in the corpus a minimum requisite number of times (five). Models were trained with the Python *gensim* package. Embedding size was set to length 300, motivated by findings that this length seems to produce robustly applicable embeddings (Landauer & Dumais, 1997; Mandera et al., 2017; Mikolov, Chen, Corrado, & Dean, 2013). Parameters within the suggested range for each model were otherwise used: window size of five, 1e-5 downsampling parameter for frequent words, and five negative samples per trial.

After models were trained, WFs and CDs for the 70,167 words were calculated within each corpus. The average need of each word was then calculated by averaging the similarity between that word and 1,000 random context words, weighted by the WFs of the context words.[2] Context words were randomly sampled (with replacement) from the set of unique words included within this study that also had an entry in the corresponding model. Random sampling of context was used to keep calculation time reasonably short; average need was originally calculated over the exhaustive set of all 70,167 context words. However, this calculation took multiple days to complete for each model. Exploratory analysis performed on the subtitles dataset suggested that random sampling of context did not substantially affect the measures of average need ($r > .99$ between average needs derived with and without random sampling of contexts). A similar subsampling procedure has been demonstrated to be effective at improving the

[2] Flat, logarithmic frequency, and squared probability weighting schemes were also tried with the Subtitles corpus. Each produced negligibly worse fits to lexical retrieval times. Worse fits were expected. When using a corpus of text as a model of the distribution of contexts of linguistic experience, and individual word tokens in the corpus as those contexts, the only sensical weighting scheme to apply is the frequency of the (context) word.

tractability of calculating a related measure, semantic diversity, without sacrificing quality of those estimates (Hoffman, Ralph, & Rogers, 2013).

Randomization models

Calculations of average need were repeated using randomization models. This was done to assess the relevance of well-structured word embeddings for the estimation of average need. A randomized version of each model was created by randomly permuting the values contained within each word embedding, prior to the calculating of average need. This process destroys any information available in a word's embedding that might encode word meaning.

Results

We start by reporting analyses of the CBOW models. All analyses were conducted on the subset of 27,056 words that (1) were used in the calculation of average needs, and (2) occurred in each of our corpora at least five times (i.e., enough for a model to produce an embedding for the word). Within this subset, only 15,994 words had BLP lexical decision data so all analyses involving BLP lexical decisions are conducted only on this set of words. Likewise, 23,767 words had ELP lexical decision data, and 23,769 had ELP word naming data.

Measures of average need were highly correlated with measures of log CD and log WF within each corpus: for log WF, values ranged between $r = .642$ ($p = 0$[3]; TASA) and $r = .858$ ($p = 0$; Wikipedia). For log CD, values ranged between $r = .622$ ($p = 0$; TASA) and $r = .849$ ($p = 0$; Wikipedia).

Very weak but reliable relationships were observed for the CBOW randomization models. For log WF, TASA $r = .016$ ($p = .01$); Wikipedia $r = .033$ ($p = 4.8\text{e-}8$); Subtitles $r = -.006$ ($p = .29$); Subtitles-90 M $r = .016$ ($p = .01$); Subtitles-180 M $r = .003$ ($p = .60$). For log CD, TASA $r = .019$ ($p = .002$); Wikipedia $r = .032$ ($p = 1.6\text{e-}07$); Subtitles $r = -.005$ ($p = .36$); Subtitles-90 M $r = .016$ ($p = .01$); Subtitles-180 M $r = .000$ ($p = .94$).

These small but reliable correlations between log CD/WF and average need in the randomization models is indicative of the fact that words could act as their own contexts. Context words were weighted by their frequency when calculating average need. Thus, common words contribute more to the weight of their own context than uncommon words. This effect is marginal and cannot account for the strong relationships seen between average need and log CD/WF in the nonrandom models. The strong relationship between average need and log CD/WF depends on the presence of well-ordered semantic representations.

CBOW average need consistently had a moderate linear relationship with measures of lexical access times: for BLP lexical decision data, values ranged from $r = -.371$ ($p = 0$; TASA) to $r = -.498$ ($p = 0$; Subtitles), for ELP lexical decision data, values ranged from $r = -.279$ ($p = 0$; TASA) to $r = -.473$ ($p = 0$; Subtitles), and for ELP naming data, values ranged from $r = -.207$ ($p = 0$; TASA) to $r = -.403$ ($p = 0$; Subtitles).

Quality of average need was also observed to depend on the size of the corpus over which word embeddings were learned. Log corpus size was predictive of the strength of relationship between estimates of average need and log WF, $r(3) = .942$, $p = .016$, and log CD, $r(3) = .959$, $p = .009$, though the pattern was not replicated for all behavioral measures: BLP lexical decision times, $r(3) = .897$, $p = .04$; ELP lexical decision times, $r(3) = .504$, $p = 0.39$; and ELP naming times, $r(3) = .359$, $p = .55$. We assume this lack of replication for some behavioral measures is simply due to a lack of power in the analyses.

Identifying that CBOW average need is strongly correlated with lexical access times establishes the presence of a linear relationship. However, it does not exclude the possibility of nonlinear relationships. Because the shape of the fit between average need and lexical access times is central to the thesis of this experiment, the possibility of nonlinear effects was considered.

CBOW average need was fit to lexical access times using polynomial regression. Linear, quadratic, and cubic terms were used. This was done in a stepwise manner. The linear term was added first. Then, the quadratic term was added to the regression. Its effect was measured, beyond the linear term. Finally, the cubic term was added and its unique effect measured. Effects for linear, quadratic, and cubic terms are displayed in Table 1. Only results for ELP lexical decision times are reported; effects for the other measures of lexical access displayed the same patterns, but with different magnitude.

It was clear from this analysis that a linear fit accounts for most of the shared variance between CBOW average needs and ELP lexical decision times. The linear term accounted for 7.79%, 20.48%, 20.67%, 22.45%, and 13.15% of the variance in lexical decision times, when considering the TASA, Subtitles-90 M, Subtitles-180 M, Subtitles, and Wikipedia corpora, respectively. In comparison, the square term only accounted for 0.33%, 0.41%, 0.52%, 0.32%, 0.35% of the variance, respectively. Likewise, the cubic term only accounted for 0.17%, 0.28%, 0.20%, 0.22%, and 0.17%. In every case, model comparison using the ANOVA function in R revealed that the unique contributions of quadratic and cubic effects were highly reliable ($p = 0$), as small as they were. This is not necessarily a problem; effects are primarily

[3] We adopt the convention of using $p = 0$ to refer to any p value that is smaller than can be accurately measured, based on the statistics software being used (i.e., 2.2e-16). This convention is adopted primarily with the goal of increasing legibility of the results section.

Table 1 The predictive validity of various measures of average need on lexical decision times taken from the English Lexicon Project

Corpus	Contribution of polynomial terms (ΔR^2 in % over previous predictors)											
	CBOW			Skip-gram			Log Frequency			Log Context Diversity		
	Linear	Square	Cube	Linear	Square	Cube	Linear	Square	Cube	Linear	Square	Cube
TASA	7.79	0.33	0.17	13.32	0.89	0.03	28.80	0.33	0.05	28.89	0.30	0.12
Subtitles-90 M	20.48	0.41	0.28	3.35	0.02	0.06	35.06	0.59	0.15	35.63	0.39	0.12
Subtitles-180 M	20.67	0.52	0.20	15.17	0.18	0.28	35.12	0.56	0.19	35.58	0.34	0.16
Subtitles	22.45	0.32	0.22	22.56	0.94	0.17	35.27	0.66	0.13	35.76	0.42	0.11
Wikipcdia	13.15	0.35	0.17	11.01	0.02	0.28	16.39	0.61	0.29	15.91	0.58	0.23

Effects from linear, quadratic, and cubic terms are reported in terms of unique variance accounted for, above previous terms. In all cases, the linear component accounts for the vast majority of variance

linear. The higher order effects could plausibly be due to distortions in estimates of average need resulting from lack of parameter optimization when training CBOW. Furthermore, the magnitudes of the quadratic and cubic effects are similar to those seen when performing polynomial regression of log WF and log CD on lexical decision times (see Table 1); comparison measures are also showing this same tendency to have slight nonlinear relationships with lexical decision times.

All three predictions were supported by results from the CBOW analyses: Average needs were (1) moderately linearly related to lexical access times, (2) strongly linearly related to the logarithms of WF and CD calculated from the same corpora, and (3) the quality of average needs depended on the quality of word embeddings, measured both by use of randomized embeddings and by using larger corpora on which to train embeddings.

We now turn our attention to the analysis of average needs derived from skip-gram models. Unlike the CBOW results, the skip-gram results did not support all three predictions of this experiment. For the smallest corpus, TASA, the observed relationship between average needs and all dependent measures was actually opposite the expected direction: $r = -.605$ ($p = 0$) for CD; $r = -.612$ ($p = 0$) for WF; $r = .362$ ($p = 0$) for BLP lexical decision time; $r = .365$ ($p = 0$) for ELP lexical decision time; $r = .300$ ($p = 0$) for ELP naming time. Models trained on the other three larger corpora showed relationships in the expected direction, but the strength of those relationships were overall quite more variable than in the case of CBOW and, when examining the three Subtitles corpus, very dependent on corpus size (see Table 1).

Skip-gram's failings with respect to the predicted results became obvious from visual inspection: The skip-gram algorithm had a tendency to overestimate the average need of uncommon words. The result is that estimated needs probability takes on a U-shaped relationship with measures of log CD/WF. Results from the full Subtitles corpus are given in Fig. 1 as an example.

We tested whether the quality of skip-gram average needs improved after taking into account the fact that the algorithm overestimates values for uncommon words. The local minima in the function describing the relationship between log CD and skip-gram average need was found for each model (i.e., for Fig. 1a, the basin of the U shape). Data were split based on whether they fell to the left or the right of this minima. Minima were estimated by first running a polynomial regression of log CD on average need (including linear, square, and cube terms), and then finding the point at which the derivative of the resulting equation equaled zero (i.e., when the function is flat, as it would be in the basin of the U shape). The log CD values where average needs were minimal were 4.982 for TASA, 4.451 for Subtitles-90 M, 4.299 for Subtitles-180 M, 4.234 for Subtitles, and 3.591 for Wikipedia. Words falling above or below these CD values will be referred to as high-diversity and low-diversity words, respectively.

When data were divided by high- and low-diversity words, skip-gram average needs consistently provided better fits to the behavioral and CD/WF data than CBOW average needs (see Table 2). However, for low-diversity words, the relationship is consistently in the wrong direction.

We considered the possibility that skip-gram may be organizing words according to their contextual informativeness. Consider the example word *cat* as it relates to the topic of *animals*. When discussing *animals*, it is very likely that knowledge of *cat* will be needed (due to their ubiquity and exemplar membership). That is to say, *cat* has high needs probability, given cues about *animals*. However, the occurrence of *cat* is not necessarily informative of a context of *animals* due to the fact that the term is used in a wide variety of contexts—for example, contexts about Dr. Seuss (the cat in the hat), jazz (cool cats), musicals (*Cats*), and the Internet (pictures of cute animals). Now consider the example of the word *capybara*. Capybaras as large rodents related to guinea pigs. Due to its relative obscurity, knowledge of capybara is seldom needed when discussing animals. However, if capybara is mentioned, there is a large chance that the context of

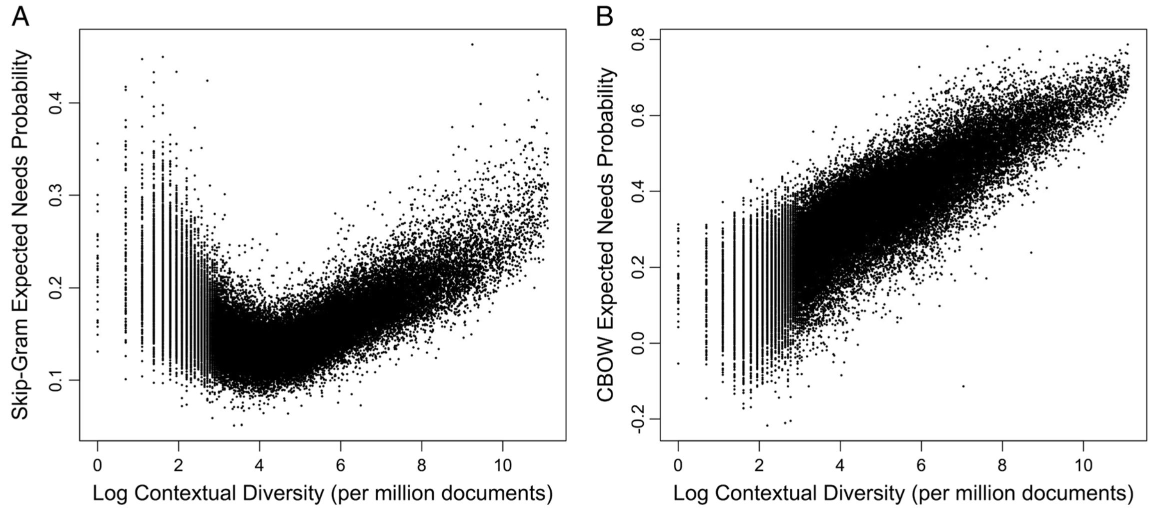

Fig. 1 The relationship between contextual diversity and **a** skip-gram average needs and **b** CBOW average needs. Results are from values derived from the Subtitles corpus

discussion is about animals (as opposed to Dr. Seuss, jazz, musicals, or the Internet). It seems plausible that skip-gram may be organizing words according to their informativeness of contexts of appearance; skip-gram learns word embeddings by predicting the identity of missing context words from a word that appeared in that context.

The information retrieval literature makes use of a measure called term frequency inverse document frequency (tf-idf). Tf-idf is measured as a word's frequency of occurrence within a specific document, multiplied by the negative log of the percentage chance it appears in any particular context (typically defined in terms of a unique document). Tf-idf measures how specific a word's use is to a particular context, weighted by its prevalence within that particular context; it is a measure of informativeness of context. Tf-idf happens to have a U-shaped relationship with log WF, much like average needs calculated from skip-gram.

Given that tf-idf has a U-shaped relationship with log WF, and that skip-gram average need estimates also had a U-shaped relationship with log WF, we tested the possibility that the two measures might be related. For each word, we calculated its average informativeness of context by averaging its tf-idf values across each document the word occurred in. We then correlated average informativeness of context with skip-gram average needs estimates. Within the Subtitles corpus, skip-gram average need and average informativeness of

Table 2 The extent to which CBOW and skip-gram average needs correlate with behavioral measures of lexical access and other operationalizations of average need

Embedding type	Hith Diwrsity Wads					Low Diversity words				
	Lexical Access Measures			Other measures		Lexical Access Measures			Other measures	
	LDRT (BLP)	LDRT (ELP)	Naming (ELP)	log CD	log WF	LDRT (BLP)	LDRT (ELP)	Naming (ELP)	log CD	log WF
Skip-Gram Embeddings										
TASA	-0.223	-0.207	-0.140	0.850	0.822	0.359	0.373	0.305	-0.797	-0.814
Subtitles-90 M	-0.397	-0.356	-0.304	0.849	0.840	0.186	0.224	0.172	-0.682	-0.695
Subtitles-180 M	-0.461	-0.416	-0.351	0.841	0.829	0.122	0.178	0.118	-0.668	-0.674
Subtitles	-0.500	-0.455	-0.386	0.827	0.815	0.056	0.120	0.072	-0.656	-0.674
Wikipedia	-0.495	-0.410	-0.335	0.856	0.839	0.009	0.074	0.068	-0.380	-0.393
CBOW Embeddings										
TASA	-0.141	-0.042	-0.019	0.569	0.578	-0.246	-0.171	-0.122	0.564	0.603
Subtitles-90 M	-0.275	-0.279	-0.256	0.719	0.757	-0.137	-0.170	-0.143	0.591	0.655
Subtitles-180 M	-0.343	-0.330	-0.288	0.718	0.756	-0.079	-0.112	-0.069	0.561	0.629
Subtitles	-0.414	-0.398	-0.343	0.770	0.802	-0.054	-0.083	-0.047	0.596	0.665
Wikipedia	-0.484	-0.412	-0.324	0.821	0.838	-0.041	-0.024	-0.014	0.460	0.565

Results are presented for five corpora and split between words with high or low contextual diversities context were reliably, but weakly, correlated ($r = -.25$, $p = 0$). Relationships of similar strengths were observed for the other corpora. Only weak evidence is available that skip-gram is organizing words according to their informativeness of context, as measured by tf-idf.

Because skip-gram average need displayed a U-shaped relationship with measures of log WF and log CD, it was anticipated that skip-gram average need would also have a nonlinear relationship with measures of lexical access. This would account for the low correlations observed between lexical access measures and skip-gram average need in some of the corpora. However, to our surprise, skip-gram average need displayed very weak evidence of having a nonlinear relationship with lexical access measures (see Table 1). The effects are, again, primarily linear, though much more variable in magnitude between corpora than CBOW average need. The variability of magnitude does seem to coincide with the degree to which skip-gram average needs are linearly related to log WF and log CD measures.

Finally, although CBOW average needs do predict measures of lexical access reasonably well (see Table 1), log WF and log CD both provide a much better fit to the behavioral data (see Table 1). So although the primary predictions of this experiment were satisfied, log WF and log CD still provide the superior quantitative estimates of average need.

Discussion

Memory retrieval is optimized to produce the most relevant items, given contextual cues (J. R. Anderson, 1991; J. R. Anderson & Milson, 1989; J, R. Anderson & Schooler, 1991). Optimization is performed over associative relationships present in the stream of experience. The current hypothesis was that CBOW and skip-gram models are solving a similar type of problem as human memory retrieval processes; they are learning to predict the presence of words in a contiguous string of text, given other linguistic cues available in that text. It was anticipated that these models would be able to replicate the main prediction of Anderson's rational analysis of memory. Namely, that estimates of needs probability derived from them would be linearly related to lexical access times. We further validated these estimates by comparing them to other measures of average need (log CD/WF).

CBOW results were consistent with predictions. These findings lend support for the hypothesis that CBOW and human memory are solving a same, or similar, computational problem of cue-memory association. Results from the skip-gram model were more complicated. Specifically, the skip-gram model appears to consistently overestimate the average need of uncommon words, resulting in a u-shaped relationship between average need derived from the models and log CD/WF.

When uncommon words were removed from analyses, the skip-gram model provided a superior fit to measures of lexical access and log CD/WF. When instead only uncommon words were analyzed, the skip-gram model also provided better fits than CBOW, but in the direction opposite what was expected. These findings lead to the conclusion that, although skip-gram is forming useful cue-memory associations, CBOW's computational description may be closer to processes underlying the optimization of human memory retrieval than the skip-gram model. CBOW better fits lexical access times across the full range of common and uncommon words. This finding was expected; the learning problem that CBOW is solving (predict a word from context) is more similar in description to needs probability than the learning problem that skip-gram is solving (predict context from a word).

What was not expected was the shape of the relationship between skip-gram average needs and lexical access times. The nonlinearity seen between skip-gram average need and log CD/WF is not translating to nonlinearity between skip-gram and lexical access times; it is merely diminishing the magnitude of the effect. It is unclear what this lack of transitivity implies about the skip-gram model and its relationships to log WF/CD and lexical access measures.

We considered the possibility that skip-gram is organizing words according to their informativeness of contexts of appearance. However, when average needs estimated from skip-gram were compared to tf-idf measures, relatively weak correlations were observed. No compelling evidence was found that skip-gram is organizing words according to context informativeness as measured by tf-idf.

We note that the results pertaining to average need are not tautological. First, in line with the second prediction of this experiment, relationships between measures of average need and the various dependent measures employed in this study were only observed in cases where nonrandom word embeddings were used. Second, consistent with the third prediction of this experiment, the quality of fits depended on the size of the corpus models were trained on. Embeddings learned from large corpora provide better fits, not because the corpora are larger per se, but because larger corpora allow for the learning of more differentiated word meanings. This effect is most clearly seen when comparing the nested subsets of the Subtitles corpus. Consider doubling the size of a corpus by exactly duplicating its content. This does not change the statistical properties of word co-occurrences and consequently should have no impact on learned embeddings. Although not described in the Results section, we verified this by comparing model a trained on a corpus constructed out of duplicating the content of the Subtitles-180 M corpus (360 million token in size, total) to models derived from (1) the full Subtitles corpus (365 million tokens in size, total) and (2) a single copy of the Subtitles-180 M corpus (180 million tokens in size, total). Average need measures were calculated from the resulting CBOW model and regressed on ELP lexical decision data using polynomial regression. The linear term for the duplicate-content Subtitles-180 M corpus accounted for 20.74% of the variance in lexical decision times—in line with the 20.67% seen for the regular Subtitles-180 M corpus, and lower than the 22.45% seen for the full Subtitles corpus (see Table 1). The marginal 0.07% difference in fit seen between the duplicate-content and regular Subtitles-180 M corpus is likely due to what effectively amounts to adding an additional training epoch over the input data. It is the statistical properties of a corpus, not its size, that determines the content of learned embeddings. All other things being equal, larger corpora will generally contain more variability of word use than smaller corpora (Baayen, 2001; Heaps, 1978). This increased variability in word use leads to the learning of better structured and more differentiated word embeddings, which are required to produce behavior consistent with predictions from J. R. Anderson's (1991) rational analysis of memory.

The reported results depend on the existence embeddings that carry information about a word's expected use. This suggests that psycholinguistic effects captured by other measures of average need (i.e., log WF/CD) are semantic in nature; effects of these variables reflect the fact that we dynamically organize semantic knowledge in accordance to its needs probability, given contextual cues present in a memory retrieval context. A similar claim has been made by others who have studied needs probability (e.g., Jones et al., 2012) and results of this nature provide a challenge for memory models that account for frequency effects, as Baayen (2010) put it, by appealing to meaningless "counters in the head" (e.g., Coltheart et al., 2001; Murray & Forster, 2004; Seidenberg & McClelland, 1989).

DSMs typically throw away information about vector magnitude; embeddings are normalized to unit length. Johns, Jones, and Recchia (2012) have argued that when embeddings encode for the various contexts of use of a word, an embedding's vector magnitude becomes informative of the semantic distinctiveness of that word. Information about semantic distinctiveness is encoded as the vector magnitude of word embeddings. Thus, something akin to average need may be explicitly represented in memory and used to prioritize access to knowledge contained within semantic memory, explaining semantic distinctiveness, WF, and CD effects.

The current results demonstrate that information about average need can also be recovered from certain classes of DSMs at the point of memory retrieval, even when information about average need is not explicitly encoded. When embeddings are learned by applying a learning objective that is consistent with the principle of likely need, e.g., CBOW, it is not necessary for average need to be encoded in memory to account for effects having to do with average need, such as CD and WF (at least, as they pertain to lexical access). The observable effects of these variables can be explained as the consequence of the distributional properties of retrieval cues interacting with the distributional properties of semantic memory during the process of memory retrieval. Such effects are a free lunch from processes of memory retrieval when semantic representations are learned by applying a learning objective that is consistent with the principle of likely need; commonly occurring words are more likely to be needed than uncommon words, given an arbitrary set of retrieval cues.

Measures of average need produced by CBOW have a linear relationship with measures of lexical access. This is a prediction that comes directly from J. R. Anderson's (1991) rational analysis of memory, and the main focus of this current research. The two other measures motivated by Anderson's concept of needs probability, contextual diversity (Adelman et al., 2006) and semantic distinctiveness (Jones et al., 2012), both have logarithmic relationships with lexical access times. Although these measures are capturing variation that is relevant for lexical access, the lack of linear relationship suggests that underlying models and motivations for these measures have a qualitative mismatch with Anderson's theory. However, the fact that CBOW correctly models the shape of this relationship (1) encourages psychologists to treat CBOW as a psychologically plausible computational model of semantics, and consequently (2) supports the argument that we may wish to conceptualize CD/WF effects, not as being explicitly encoded in memory structures, but as being the byproduct of retrieval cues interacting with memory structures that are shaped to be optimally accessible with respect to needs probability over an historic distribution of memory retrieval contexts.

As a final point, CBOW average needs had a linear relationship with lexical access times and that fit was moderate in magnitude. However, the amount of variance accounted for was still substantially less than log CD/WF (see Table 1). Thus the possibility arises that, although it does capture the qualitative effect that was expected, quantitatively CBOW may be a poor model of memory retrieval. In the next experiment, we argue the reason for this is that our calculation of average need treats each dimension in an embedding vector as equally relevant and this is an unreasonable constraint on model testing.

Experiment 2

Word embeddings specify the dimensions along which the uses of words are most clearly distinguished. Conceptually, embeddings are feature vectors that represent a word's meaning. However, these feature vectors are learned from experience without supervision rather than being specified by behavioral research findings (e.g., McRae et al., 2005) or pseudorandomly generated (e.g., Hintzman, 1988). Interestingly, even though embedding methods are never instructed on what types of features to learn, what they do learn is consistent with the primary dimensions over which humans make semantic distinctions: valence, arousal, dominance, concreteness, animacy. This has been demonstrated for the skip-gram model (Hollis & Westbury, 2016; Hollis, Westbury, & Lefsrud, 2016) as well as CBOW (Hollis & Westbury, 2017). Both models are honing in on, to some extent, psychologically plausible semantic spaces in an unsupervised manner.

It is useful to think about context of memory retrieval as the semantic features underlying probe stimuli (e.g., Howard & Kahana, 2002). It is not the word *dog* that provides context as a probe stimuli. Rather, it is the underlying meaning of *dog* that provides context. Two physically dissimilar stimuli (e.g., dog; schnauzer) act very similarly as context because they are united by their underlying semantics.

In a multicue retrieval context, not all cues will be equally salient; some will most certainly be attended to more or less than others, depending on the task or goal. Indeed, individual features learned by the skip-gram model vary substantially in how much they contribute to the prediction of forward association strengths (Hollis & Westbury, 2016). Since not all features are equally relevant in any particular task context, we should expect the ability of embedding models in predicting memory retrieval effects to depend on whether the contributions of individual semantic features are equally weighted or left free to vary.

In Experiment 1, needs probability was measured via cosine similarity. One of the limitations of cosine similarity is that it weights each feature equally; as a means of estimating needs probability, cosine similarity treats each available semantic feature as if it were equally relevant to the task being modeled. Consequently, the results of Experiment 1 are likely an underestimate of how well CBOW and skip-gram embeddings fit lexical access data.

The purpose of Experiment 2 was to provide an assessment of how well CBOW and skip-gram embeddings fit behavioral measures of lexical access when the contributions of each embedding dimension could have independently weighted contributions.

Method

Experiment 2 used the same words, embeddings, and dependent measures as Experiment 1. However, average needs were not calculated according to the method of Experiment 1. Instead, we took a regression approach. Embedding dimensions were regressed on dependent measures as a means of assessing the extent to which variation within a particular embedding dimension (i.e., a semantic feature) could account for variation within the dependent measures. This method accommodates for the fact that semantic features may differ in terms of salience or relevance for any particular task and thus may contribute to processes of memory retrieval by varying degrees.

Results

For ease of readability, we only report results for ELP lexical decision data. Patterns of results for the BLP lexical decision data and ELP naming data were similar but differed in the amount of variance that could be accounted for (BLP lexical decision data are generally most predictable, followed by ELP lexical decision data and then ELP naming data). We likewise limit written results to the Subtitles corpus for ease of readability. Summary results for other corpora are presented in tabular form. All results from regression analyses are after model validation using k-fold cross validation (k = 10). All regressions were performed over the same $n = 27{,}056$ used in Experiment 1, unless otherwise stated.

Predictive validity of CBOW and skip-gram embeddings

Individual dimensions from CBOW and skip-gram word embeddings (length 300 each) were regressed on ELP lexical decision times as a means of assessing the relevance of each feature to the task of lexical decision.

CBOW embeddings from the Subtitles corpus accounted for 40.47% of the variance in ELP lexical decision times. The value for the CBOW randomization model was much lower: 0.14%. In contrast, CBOW average needs from Experiment 1 only accounted for 22.37% of the variance in ELP lexical decision times. Leaving each dimension of the model free to contribute independently substantially improves the fit of the model. Readers are reminded that all results from regression analyses are after model validation using k-fold cross validation (k = 10), so increases in fit are not due to simply having a regression equation with more degrees of freedom.

To our surprise, skip-gram embeddings derived from the Subtitles corpus also accounted for a large portion of the variance in ELP lexical decision times: 37.52%. The value for the skip-gram randomization model was comparable to the CBOW randomization model: 0.22%. The skip-gram average needs from Experiment 1 accounted for 22.56% of the variance.

We performed a series of regressions to see how much variance CBOW embeddings accounted for beyond skip-gram embeddings and vice versa. Skip-gram embeddings accounted for 4.33% unique variance over CBOW embeddings, whereas CBOW embeddings accounted for 7.29% unique variance over skip-gram embeddings. Despite using converse learning objectives, the two models are honing in on largely the same variance, as it pertains to lexical decision.

An observation from Experiment 1 was that the skip-gram model overestimates the likely need of uncommon words. When words were split according to common and uncommon words, skip-gram provided a better fit to lexical decision times that CBOW on both halves, but in the wrong direction for uncommon words. This led to the conclusion that skip-gram average needs were capturing meaningful variation in lexical decision times, but for perhaps incidental reasons; the model does not display the correct qualitative pattern across the full range of words.

For the current experiments, we limited analyses to only those words that occurred five times or more in each of the five corpora. This necessarily biases analyses away from very uncommon words, which is not ideal because those are the words that are most highly informative for comparing CBOW and skip-gram as models of semantic memory. We thus reran the above regressions using a larger set ($n = 35,783$) of words. This set composed the intersection between (1) all words that ELP lexical decision data were available for, and (2) all words where CBOW and skip-gram embeddings were available for in the Subtitles corpus. This primarily included very uncommon words and shifted the median word frequency from 2.69 per million (982 occurrences) to 1.52 per million (545 occurrences).

CBOW embeddings accounted for more variance in lexical decision times when the larger word set was used: 41.90% versus 40.47%. In contrast, skip-gram embeddings accounted for less variance when the larger word set was used: 29.36% versus 37.52%. We again tested how much variance CBOW and skip-gram embeddings accounted for, above the other. CBOW accounted for 16.38% of the variance over skip-gram embeddings, whereas skip-gram only accounted for 3.84% unique variance above CBOW embeddings. When effects of length and orthographic neighborhood size were first accounted for, these numbers dropped to 9.14% for CBOW and 2.21% for skip-gram. Over a larger set of words that is more inclusive of very uncommon words, CBOW substantially outperforms the skip-gram model at accounting for lexical access times.

Consistent with analyses from Experiment 1, CBOW provided a much better fit to lexical access data than skip-gram across the full range of common and uncommon words. These effects are not spurious and instead have to do with the information contained within embeddings, as evidenced by the poor performance of randomization models.

Comparing CBOW embeddings to CD and WF

For comparison, log WF calculated from the Subtitles corpus accounted for 35.27% of the variance in ELP lexical decision times and log CD accounted for 35.76% of the variance. This is lower than the 40.47% accounted for by CBOW embeddings.

We predicted that CBOW embeddings should account for little to no variance in lexical access measures above and beyond WF and CD on the thesis that embeddings are feature vectors of word meaning, and that CD and WF are measures of how likely needed a word's meaning is, aggregated over numerous semantic features about that word's underlying meaning. Both sources are capturing the effects of many varied semantic sources on lexical access. Likewise, it was expected that neither CD nor WF would account for much variance in lexical access measures beyond that accounted for by CBOW embeddings.

Word length and orthographic neighborhood size are also both measures thought to play an important role in lexical decision. Indeed, a common benchmark for assessing the relevance of measures thought to play a role in word processing is to see if they account for any variance in behavioral measures above and beyond that accounted for by word length, orthographic neighborhood size, and log WF (Adelman, Marquis, Sabatos-DeVito, & Estes, 2013).

In terms of measuring the predictive validity of CBOW embeddings, it is pertinent to test if they account for any variation beyond length and orthographic neighborhood size. The difficulty is, when performing a direct comparison of the ability of CBOW embeddings, WF, and CD to account for unique variance in lexical decision, also including word length and orthographic neighborhood size may unfairly favor CBOW embeddings. It is known that CD and WF are both correlated with length and orthographic neighborhood size, however we have not yet tested the relationship between these orthographic measures and CBOW embeddings; some of the variance that would be accounted for by WF and CD is instead being accounted for by length and orthographic neighborhood size. To accommodate for this possibility, we perform two sets of comparisons between CBOW, WF, and CD: one set including length and orthographic neighborhood size and one set not including these variables. In both sets of analyses, the goal was to identify how much variance each of CBOW, WF, and CD accounts for in lexical decision times above other measures, as well as how much variance each measure accounted for on their own. Results are presented in Table 3.

These model comparisons produced some interesting patterns of results. First, it was very clear that CBOW was much more sensitive to corpus size than WF and CD. When trained on the TASA corpus, CBOW performed relatively poorly at predicting LDRT effects (15.78% variance accounted for by itself), compared to CD and WF (28.89% and 28.80%). However, as corpus size increased, CBOW had consistent performance gains. This can most clearly be seen when looking across the nested subsets of the Subtitles corpus: 35.35% variance of Subtitles-90 M, 38.75% variance on Subtitles-180 M, and 40.47% on the full Subtitles corpus. In contrast, neither CD nor WF showed substantial gains in predictive validity across these three corpora (CD 35.63%, 35.58%, 35.76%; WF 35.06%, 35.12%, 35.27%). These results are consistent with previous findings. Brysbaert and New (2009) found that the predictive validity of word frequency

Table 3 The unique contributions of contextual diversity (CD), word frequency (WF), and CBOW embeddings (CBOW) in predicting lexical access times (English Lexicon Project lexical decision times)

Analysis	Unique effect (ΔR^2 in % over other predictors)									
	Analysis Excludes Len + ON					Analysis Includes Len + ON				
	TASA	Subt-90 M	Subt-180 M	Subtitles	Wikipedia	TASA	Subt-90 M	Subt-180 M	Subtitles	Wikipedia
CBOW	15.78	35.35	38.75	40.47	35.83	10.51	18.58	20.88	22.11	20.01
CD	28.89	35.63	35.58	35.76	15.91	18.55	19.65	19.65	19.73	14.42
WF	28.80	35.06	35.12	35.27	16.39	17.78	18.30	18.36	18.41	14.08
CBOW (after CD)	8.02	6.76	7.80	8.76	26.12	2.55	3.49	4.08	4.79	9.89
CBOW (after WF)	7.90	6.73	7.88	8.92	25.79	2.61	3.92	4.70	5.56	10.01
CD (after CBOW)	21.13	7.04	4.63	4.05	6.19	10.58	4.55	2.85	2.42	4.30
CD (after WF)	0.23	0.58	0.47	0.50	0.20	0.81	1.85	1.83	1.93	0.43
WF (after CBOW)	20.92	6.44	425	3.72	6.35	9.88	3.64	2.19	1.87	4.08
WF (after CD)	0.14	0.01	0.01	0.01	0.68	0.04	0.50	0.55	0.61	0.10
CBOW (after CD + WF)	7.97	6.76	7.79	8.76	25.59	2.58	3.25	3.79	4.42	9.84
CD (after CBOW + WF)	0.29	0.60	0.38	0.34	0.00	0.78	1.18	0.92	0.78	0.27
WF (after CBOW + CD)	0.08	0.00	0.00	0.00	0.16	0.07	0.27	0.26	0.23	0.05

Data are reported for measures derived from five corpora. Unique contributions are presented both with and without first accounting for variance attributable to worth length (Len) and orthographic neighborhood size (ON)

measures quickly plateau in quality after 20 million tokens. However, the quality of CBOW embeddings, as measured on a variety of different tasks, continues to improve well into billions of tokens. CBOW does, however, perform quite poorly on small corpora where insufficient distributional information is available to adequately learn vector representations for words (Lai, Liu, Xu, & Zhao, 2016). To understand this difference in performance between CD, WF, and CBOW, it is important to consider the differences in the ranges of variability between word frequency and co-occurrence frequency: there is far more variability in the distribution of co-occurrences than word frequencies because word frequency is determined by words whereas co-occurrence is determined by Word × Word relationships; with large numbers of elements, there are always vastly more possible combinations of elements than elements themselves. It makes sense for any sort of measure derived from frequencies to arrive at a stable estimate before a measure derived from co-occurrences. It likewise makes sense for co-occurrence measures to perform much poorer on small data sets for the same reasons; for small corpora, co-occurrence information is too sparse despite individual words possibly having occurred numerous times.

A second interesting pattern of results is that CBOW accounts for more unique variance than either WF or CD on the three largest corpora (i.e., Subtitles-180 M, Subtitles, Wikipedia). This is most pronounced on the Wikipedia corpus where CBOW accounts for 25.59% of the variance in lexical access measures above CD and WF (9.84% after accounting for length and orthographic neighborhood size), whereas CD and WF account for 0.00% and 0.16%, above the other two variables, respectively (0.27% and 0.05% after length and orthographic neighborhood size). This is perhaps an unfair comparison, given that CD and WF are effectively the same measure on this corpus and over these $n = 27{,}056$ words ($r > .99$). But even when looking at these variables in a pairwise fashion, the differences in unique variance is very large: 26.12% and 25.79% for CBOW over CD and WF respectively (9.89% and 10.01% after length and orthographic neighborhood size) compared to 6.19% and 6.35% for CD and WF over CBOW respectively (4.30% and 4.08% after length and orthographic neighborhood size).

A third interesting pattern is that as corpus size grows, less unique variance is attributable to CD and WF and more to CBOW. So it is not just that CBOW is consistently capturing more of the variance that would otherwise be attributable to CD and WF as corpus size increases (less unique variance attributable to CD and WF). In large corpora, CBOW is also picking up on additional information that CD and WF do not carry (more unique variance attributable to CBOW). This pattern is clearly seen when looking at the unique variance attributable to each variable on the two-variable comparisons (e.g., CBOW after CD vs. CD after CBOW) over the three Subtitles corpora. The pattern is present whether or not effects due to length and orthographic neighborhood size are accounted for first.

As a final observation, it bears noting that although CBOW and the other two variables have quite a bit of shared variance, CBOW never completely excludes the other two, nor vice versa. Focusing specifically on pairwise comparisons over the Subtitles corpora, where none of CBOW, CD and WF are at large deficits (unlike TASA, which is likely too small of a corpus for CBOW to be effectively trained on, and Wikipedia where CD and WF measures are exceptionally poor), the amount of unique variance attributable to CBOW ranges from 6.73% to 8.92% (3.49% to 4.79% after length and orthographic neighborhood size), versus 4.05% to 7.04% for CD (2.42% to 4.55% after length and orthographic neighborhood size), and 3.72% to 6.44% for WF (1.87 to 3.64% after length and orthographic neighborhood size).

Discussion

The results of Experiment 2 provide support for two conclusions. First, CBOW has more psychological relevance than skip-gram as a model of word meaning. This is evidenced by the fact that CBOW embeddings account for substantially more variance in lexical access times than skip-gram embeddings when considered across the full range of common and uncommon words.

Second, the effects that CD and WF have on lexical access are largely semantic in nature. This is supported by that fact that large portions of the variance in lexical access times accounted for by CD and WF can also be accounted for by embeddings produced by CBOW, which are semantic feature vectors. These results helps reconcile the fact that so few semantic features have been found that impinge strongly on tasks that require access to word meaning, above and beyond WF. We adopt J. R. Anderson's (1991) stance that the primary consideration of memory retrieval is needs probability. Needs probability is definitionally dependent on semantics. CD and WF are indexing the cumulative effects of numerous semantic factors that influence the need for knowledge across a broad range of contexts. From this theoretical stance, the question of where variation in retrieval times is coming from should actually be approached from a direction that is opposite the typical one: How much variance does WF account for in lexical decision times beyond semantic features? These results suggest that the answer is, at most, very little.

Third, in terms of operationalizations of average need, CBOW embeddings provide the superior account when used with large corpora, whereas CD and WF appear better suited for smaller corpora (less than ~90 million tokens). However, in no cases did CBOW embeddings exclude CD or WF from predicting lexical decision times, or vice versa. It remains unclear whether this is because the measures are actually picking up on unique variation, have differential dependency on corpus size (possibly, CBOW might account for all of the relevant variation with a sufficiently large corpus), or due to lack of parameter optimization of CBOW. All of these options should be pursued in future research. At the current moment, however, we can confidently conclude that CBOW embeddings, WF, and CD all have a large degree of shared variance between them insofar as their ability to account for lexical decision times is concerned. This is to be expected under the thesis that all three measures get their predictive power via their shared relationship to the construct of needs probability.

General discussion

The reported research presents multiple findings and theoretical implications that are of relevance to the study of language and memory.

Differences between CBOW and skip-gram

CBOW and skip-gram differ in the predictive validity of lexical access times, with CBOW providing the superior fit across the full range of common and uncommon words in all cases. CBOW average needs also provided the expected linear fit to logarithmic CD values, whereas those of the skip-gram model displayed a U-shaped relationship. Furthermore, CBOW lent itself to psychological interpretation more readily than skip-gram. Overall, of the two models, CBOW appears to be the one that will be of most interest to psychologists; its behavior and formal learning objective are both consistent with J. R. Anderson's rational analysis of memory (1991; J. R. Anderson & Milson 1989; J. R. Anderson & Schooler, 1991).

It bears noting that of the two models, skip-gram generally produces more desirable behavior on engineering problems, particularly when very large corpora (multiple millions of tokens) are used to learn embeddings (e.g., Mikolov, Chen, et al., 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). However, it would be fallacious to claim that higher accuracy on complex tasks necessarily means one model is more in line with actual human performance than another model. Humans are rarely perfectly accurate on anything, and correctly modeling errors is often stronger evidence for a model than correctly modeling nonerrors. Thus, for any task, further gains past some performance threshold start becoming strongly indicative of very unhumanlike behavior.

Experiment 1 demonstrated that, within skip-gram, words with the highest global similarity to all other words are those that are either very frequent or very infrequent. If '*knife*' is a probe word, then of the near-synonyms of *cut*, *slash*, and *lacerate*, the high-frequency *cut* and the low-frequency *lacerate* are expected to be more accessible than the middling-frequency *slash*. As model benchmarking tests are made more difficult (i.e., by drawing on more obscure, less frequent, test items), skip-gram will necessarily outperform CBOW due to how the two models organize their entries;

low-frequency words in skip-gram are, on average, more similar to an arbitrary probe than they are in CBOW (Experiment 1) and are thus more accessible on average.

Skip-gram prioritizes retrieval of information that is as general as possible (high frequency) or as specific as possible (low frequency), downweighting the middle range. For applied problems like returning results from a search engine query, retrieved documents would be optimally useful for users who either want very general, entry-level information or very specific and focused knowledge, ignoring the middle. At face value, this appears to be a desirable property for many information retrieval systems, in contrast to CBOW where very specific information would be difficult to access. So although skip-gram may be better solving an interesting information retrieval problem, it may still not be a more useful model of human semantic memory than CBOW.

Word frequency and contextual diversity effects

The results of both Experiment 1 and Experiment 2 bolster previous claims that WF and CD effects in language processing tasks are likely semantic in nature (e.g., Jones et al., 2012). Experiment 1 motivates this conclusion by demonstrating that average needs calculated from CBOW are strongly correlated with both log WF and log CD. Experiment 2 motivates this conclusion by demonstrating that a large portion of the variability in lexical access measures accounted for by CD and WF can also be accounted for by the semantic features comprising CBOW embeddings; when considering the Subtitles corpus (i.e., the corpus that produces CD and WF measures with the most predictive validity), CD alone accounts for 35.76% of the variance in lexical access times. However, only 2.42% is attributable uniquely to CD, above and beyond the effects of CBOW embeddings, word length, and orthographic neighborhood size. For WF, this number is 1.87%. These results are difficult to reconcile for models that explain CD or WF effects in terms of mechanisms that are indifferent to the semantic content of memory.

It is possible that the 2.42% of variance attributed to Subtitles-based CD and 1.87% to WF per se reported here is an overestimate. Experiments 1 and 2 both identified a trend that as corpus size grew, the amount of variance uniquely attributable to CD and WF was diminished. It is a fair question whether or not this trend continues if larger corpora are employed. This is in addition to the facts that (1) no parameter optimization was performed for CBOW, and (2) it is difficult to know how much one is distorting the relevant co-occurrence structure of texts when preprocessing it to remove nonalphabetical characters: a problem that has no consequence for the calculation of WF and CD but most certainly has consequences for models that learn word meanings from co-occurrence structure.

An objection to the above argument is that humans are unlikely to have much more linguistic exposure than what the models in the currently reported experiments have received. Thus, there is little psychological relevance in considering the effects of corpora larger than are used in the current research. Brysbaert, Stevens, Mandera, and Keuleers (2016) point out that the theoretical maximum linguistic exposure of an individual per year is quite limited. Estimates range between 11.688 million tokens/year (based on the distributions of social interactions and speech in those interactions), 27.26 million tokens/year (watching television 24 hours per day), and 105 million tokens/year (reading 16 hours per day). Thus, a human will have been exposed to, at maximum, between 220 million and 2 billion words by the time they are 20 years old, depending on the source by which linguistic exposure is estimated. It is very likely that corpora used in the current research already exceed the realistic linguistic exposure of most university undergraduates.

Humans are limited in the amount of linguistic exposure they can feasibly receive in a lifetime; that is a brute fact of our finite existence. However, such an objection to exploring effects within much larger corpora neglects differences between humans and DSMs. DSMs typically receive only linguistic input whereas humans additionally have a large amount of extralinguistic experience. This extralinguistic experience is most certainly leveraged during language learning (Landauer, 2002). It is not reasonable to compare a 20-year-old human with 220 million words worth of experience to a DSM that also has 220 million words worth of experience; the human's linguistic knowledge will be much richer because it is supported by a multisensory, high-dimensional array of extralinguistic input.

There are at least two useful approaches for attempting to make an equivalence between DSM experience and human experience: (1) improve the quality of input to DSMs by additionally providing them with extralinguistic information, or (2) improve the quantity of input to DSMs to make up for the fact that DSMs receive a much more impoverished learning input than humans. It is thus reasonable to extrapolate the performance of DSMs to corpus sizes that exceed human linguistic exposure; the rationale is that increased quantity of experience is standing in for the quality of experience that DSMs lack; not an ideal solution, but still a reasonably valid one. However, we should consider at what point the missing quality of human experience has been adequately matched by the increased quantity of DSM experience. The author is aware of no work that attempts to address the quantity of relevant information DSMs miss out on, in comparison to humans, by being limited to only linguistic input.

Both Experiments 1 and 2 used randomization models to verify that observed effects depended on the presence of well-ordered word embeddings. No effects were replicated with randomization models. This finding illustrates that some effects of memory retrieval can be accounted for by the relationship between retrieval cues and the structure of the encoded contents of memory. Thus, memory models that utilize random vector representations run the risk of producing over-complicated accounts of memory, because they have to posit extra functional details in memory that would otherwise be accounted for by the structure of entities in memory as that structure relates to retrieval cues. Similar points has been made by others working with DSMs (e.g., Johns & Jones, 2010; Jones & Mewhort, 2007). The results of these experiments suggest that some types of effects, like WF and CD effects, may come for free during memory retrieval with an appropriate representational structure for semantic memory.

The shape of count effects

Generally, logarithmic transformations are applied to count data (WF, CD) when predicting lexical access measures with regression methods. Although this started off "merely" as an analytical convenience (for WF, see Howes & Solomon, 1951) the specific shape of a count variable's effect comes with theoretical implications. Consider Murray and Forster's (2004) serial position model. They argue that lexical retrieval involves a process of serial iteration over a list of candidates, ordered by occurrence frequency. This comes with a theoretical commitment that the "correct" form of the frequency effect is rank order. Generally, however, word frequency effects are conceptualized as learning effects (see, e.g., the principle of repetition vs. the principle of likely need in Adelman et al., 2006; Jones et al., 2012). When considering learning,

> if the frequency of a word is treated as an index of the amount of practice with that word, then, as with any skilled performance, improvement as a function of practice must approach a limit, the so-called irreducible minimum and so it would be expected that the effects of practice would gradually diminish in size. (Murray & Forster, 2004).

The established learning rules we use in psychology (i.e., the delta rule, the Rescorla-Wagner rule; Rescorla & Wagner, 1972) are rules of proportional change, which means that learning would best be described by an exponential function. This would result in the logarithm of practice attempts being linearly related to performance measures. Stated alternatively, a logarithmic effect for any sort of count measure is expected from proportional learning rules when the quantity of practice with a stimulus is determined by that count measure.

A challenge to the above stance is that effects of learning seem to be better described by a power function (the so-called power law of learning) than by an exponential function. Whereas the logarithm of practice should be linearly related to performance under an exponential learning function, performance and practice are instead expected to be linearly related on a log-log scale under a power function. Heathcote, Brown, and Mewhort (2000) demonstrated that when data from multiple participants are averaged together, the average data are best fit by a power function. However, when participant data are fit individually, exponential functions provide the superior fit. These results suggest that the apparent power-law of learning is an artifact of averaging. R. B. Anderson and Tweney (1997) have likewise demonstrated that forgetting rates over averaged data can artifactually appear to follow a power law when, in fact, the actual form for individual participants is exponential. These results are consistent with the claim that learning is best described by an exponential function, not a power function.

Questions about the actual shape of learning effects aside, there are empirical findings that suggest frequency effects are not learning effects in the first place. One of the motivations for contextual diversity is the finding that repeated exposures to a word do not facilitate that word's later recall unless those repetitions are accompanied by a change in context (Adelman et al., 2006). Supporting evidence has also been reported by Jones et al. (2012), who find that in a pseudolexical decision task to recently encountered nonwords, recognition is only facilitated by number of exposures if those exposures occur across multiple contexts. If the stimulus is repeated in the exact same context, repetition does not facilitate recognition. This is difficult to reconcile for theories that suppose frequency effects reflect learning due to exposure to a stimulus.

Murray and Forster (2004) point out that, if frequency effects primarily reflect processes of learning, then as people accrue more exposure to language, frequency effects should flatten off (as people acquire near-asymptotic skill with a greater variety of words). Thus, we should see smaller frequency effects for older readers. However, this is empirically not the case (e.g., Tainturier, Tremblay, & Lecours, 1989). These findings are reconciled by Murray and Forster's (2004) model by explaining that frequency effects reflect the relative ordering (and hence accessibility) of items in memory rather than absolute amounts of learning. Under the model of Murray and Forster, the locus of frequency effects is within processes of memory retrieval, not processes of learning.

Operationalizations of needs probability and average need come with different theoretical commitments to the shape of count effects than do learning theories. If count effects are being informed by Anderson's rational analysis of memory, it is expected that they will demonstrate a linear relationship with retrieval times. Yet neither CD, WF, nor SD_Count have this linear relationship with lexical retrieval times. Rather, they are all logarithmically related to such measures. The constructs, as they relate to lexical retrieval times, seem to be more consistent with learning theories than memory retrieval theories.

Whereas Adelman et al. (2006) is nonspecific about the underlying mechanisms of CD effects, Jones et al. (2012) provide a computational model that encodes semantic distinctiveness as the vector magnitude of semantic representations. Since Jones et al. (2012) are making a specific claim to a model, but the model does not produce the expected linear effect, that model needs to be revisited in terms of its relationship to the concepts of needs probability and average need. In contrast, the CBOW model does produce the expected linear effect of average need. This is most likely due to the fact that its learning objective is consistent with Anderson's concept of needs probability.

One of the important differences between Jones et al.'s (2012) semantic distinctiveness model and CBOW is how they come to their operationalization of average need. Within the semantic distinctiveness model, average need is explicitly encoded in memory structures in the form of vector magnitude. In CBOW, average need is not explicitly encoded. Its effects emerge as a byproduct of the process of memory retrieval over semantic representations acquired through the learning objective of predicting the probability of a word's occurrence conditioned on its context.

We have argued that the results of these experiments support the claim that word frequency effects are best interpreted as being effects of memory retrieval, not learning, vis-à-vis their relationship to operationalizations of average need. This is a claim that is consistent with Anderson's rational analysis of memory and supported by the current empirical results. However, this claim does create a gap in our general understanding of the shape of word frequency effects. Under theories that emphasize frequency effects as learning effects, there is a clear explanation as to why word frequency effects are logarithmic: that is the shape expected from proportional learning rules. Under a theory that emphasizes memory retrieval, that explanation is invalid. However, the invalidity of this explanation does not change the fact that, empirically, word frequency is very well fit to lexical retrieval times by a logarithmic transformation. For a memory retrieval interpretation to carry its weight, it must also provide a nonlearning-based account of the particular shape of word frequency effects.

One possible explanation of the shape of the word frequency effect comes from the psychophysics of Fechner (1860/1965). Consider the case of the poor university student (a situation that many readers will have firsthand experience of). To the poor university student, a $10 bill has high subjective value. It is food for a day, or possibly a few days, if one has a predilection toward ramen noodles. That same $10 bill has much lower subjective value to the billionaire tech mogul. In fact, it might have negative subjective value; the act of taking out one's wallet and finding a fold to place the bill into may actually have an opportunity cost much greater than $10 if that time could have otherwise been spent making money through more effective means.

The core insight in the above example, as identified by Fechner, is that the subjective value of an external stimulus must be determined in relation to what a person already possesses. A $10 bill has more subjective value to the starving university student than to the tech mogul because $10 is proportionally more of the student's savings than the mogul's. This proportionality of value means that objective quantity ($10 of currency) must scale logarithmically with subjective value (what $10 means to a specific person). Logarithmic effects are thus expected to be present any time there is an influence of an objective stimulus quantity (e.g., light intensity) on a behavior or verbal report that requires a judgment of value (e.g., brightness) is measured.

Consider that the concept of needs probability is a statement of value: How useful will this memory trace or lexical entry be to me in the near future? Any attempt to accurately capture the phenomenon in terms of quantities external to the observer (i.e., CD, WF) should result in a logarithmic relationship with the subjective value if those external quantities in fact inform the subjective value. By extension, any mental process that depends linearly on that subjective value (i.e., as lexical retrieval depends linearly on needs probability) should also be logarithmically related to the external quantity. The reason why measures of average need derived from CBOW do not have a logarithmic relationship with lexical retrieval times is because they are determined by memorial representations internal, not external, to the observer (in this case, the observer being a model). Those memorial representations are the linear, additive mental basis of subjective value (CBOW embeddings do in fact demonstrate linear, additive properties: Mikolov, Yih, & Zweig, 2013).

An implication of this line of reason is that it is conceptually incorrect to think of CD, and by extension WF, literally as measures of average need. They are no more measures of average need than light intensity is of brightness. Rather, they are two sides of a phenomenon, one side internal, relational, and subjective, and the other side external, absolute, and objective. Lawfully related to be sure but, in the spirit of Fechner, conceptually distinct entities that are parts of separate phenomenal worlds.

Future directions

The current findings help make sense of an observation from the literature on connected text reading. Despite having a large effect on the reading times for individual words, WF has very little influence on reading times for larger units of text like sentences and stories (e.g., Kuperman, Drieghe, Keuleers, & Brysbaert, 2013; Wallot, 2014; Wallot, Hollis, & van Rooij, 2013). In lexical decision and naming tasks, there exist no stable contextual cues in the task environment to aid with memory retrieval. The absence of stable contextual cues aligns needs probability with measures of how often a word occurs; absent information about context, the best guesses as to which words will be needed are exactly the words that occur most often across contexts. However, as the number of available contextual cues accumulates (e.g., as one reads through a news article or a novel) the needs probabilities of words become more clearly specified by that context, meaning the general case carries less predictive validity. Norris (2006) has made a claim to the same effect when discussing optimal behavior in word recognition: "The better the perceptual evidence, the smaller will be the influence of frequency. Frequency can never override reliable perceptual evidence" (p. 331). This line of reason leads to at least two predictions: one, when reading connected texts, WF (and CD) effects should be more pronounced at the beginning of the text than at the end of the text. Two, the speed of reading a word during connected text reading should primarily be determined by how well it is predicted by the cues available in previous parts of the document. Such evidence would suggest that WF and CD are incidental to memory retrieval, not causally involved in it; WF and CD are constructs entirely external to the observer; no "counters in the head" tracking them exist (see also, Baayen, 2010). However, they are nonetheless lawfully related to the actual causal, process-based construct of needs probability.

The presented research suggests clear avenues for future theoretical work on connected text reading. Options for methodological improvements to calculating needs probabilities also present themselves. One of the interesting properties of embeddings produced by both CBOW and skip-gram is that they have additive properties. For example, the embedding for *king* minus that of *man*, plus that of *woman* approximates the embedding for *queen* (e.g., Mikolov, Yih, & Zweig, 2013). An embedding for a document can be created by summing the embeddings of words it contains. A line for future research is to test whether average needs derived from CBOW and skip-gram can be improved upon by using document embeddings rather than word embeddings as a means of quantifying context.

Prudence is warranted before attempting to generalize the results of these experiments too broadly. Our models were all constructed with default parameter settings. Both CBOW and skip-gram have numerous parameter values, all of which can impact the performance of these models (e.g., Mandera et al, 2017; Mikolov, Chen, et al., 2013). It will be worth the time for future research to systematically explore the parameter space of these models to examine how robust their observed behaviors are. The two parameters of primary interest are window size and vector length. Window size (how many words before or after the target word are used when learning embeddings) is central to the model's definition of context when optimizing over its learning objective. Thus, we should expect these models to produce behavior more or less consistent with observed human performance as the model's definition of context converges on or diverges away from psychologically plausible context sizes. Vector length determines how many semantic features compose a word's embedding. Using too few dimensions means these models produce overly general representations of word meaning, and too many dimensions means these models produce overly specific representations of word meaning. There are specific scales over which semantics tend to be considered by humans (e.g., Rosch, 1975); vector dimensionality will play a central role in producing a psychologically plausible representation of meaning.

Conclusion

The described research was originally motivated by a desire to understand why CBOW and skip-gram models display such markedly better performance than other DSMs at capturing the semantic relationships between words. A possibility became apparent in a study of Anderson's rational analysis of memory: the computational problems being solved by CBOW and skip-gram are both descriptively similar to Anderson's theory of the function of memory. However, CBOW's learning objective (to be able to predict a word's occurrence from context information) is descriptively and formally a better fit for needs probability than skip-gram's learning objective (to be able to predict context of use from a word). In line with this difference, CBOW produces behavior more consistent with predictions from Anderson's rational account of memory than skip-gram. We are left with a possible answer to our original question in the case of CBOW: It accurately captures semantic relationships between words because it is solving a similar computational problem as human memory. There has been a surprising dearth of psychological research directed towards understanding the CBOW and skip-gram models of lexical semantics. The current results suggest CBOW, in particular, warrants study as a psychologically plausible model of meaning.

Acknowledgments The author thanks Marc Brysbaert, Brendan Johns, and an anonymous reviewer for helpful comments offered on a previous draft of this manuscript.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814–823.

Adelman, J. S., Marquis, S. J., Sabatos-DeVito, M. G., & Estes, Z. (2013). The unexplained nature of reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1037–1053. doi:10.1037/a0031829

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14*(03), 471–485.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96*(4), 703.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*(6), 396–408.

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition, 25*(5), 724–730.

Baayen, R. H. (2001). *Word frequency distributions* (Vol. 18). New York: Springer Science & Business Media.

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon, 5*(3), 436–461.

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

Baroni, M., Dinu, G., & Kruszewski, G. (2014, June). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1: Long Papers, pp. 238–247). doi:10.3115/v1/P14-1023

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). *Enriching word vectors with subword information*. Retrieved from https://arxiv.org/abs/1607.04606

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.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016, July 29). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology, 7*. doi:10.3389/fpsyg.2016.01116

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(03), 181–204.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204.

Durda, K., & Buchanan, L. (2008). Windsors: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods, 40,* 705–712. doi:10.3758/BRM.40.3.705

Fechner, G. (1965). *Elemente der Psychophysik [Elements of psychophysics]*. New York: Holt, Rinehart & Winston (Original work published 1860).

Heaps, H. S. (1978). *Information retrieval: Computational and theoretical aspects*. New York: Academic Press.

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*(2), 185–207.

Herdağdelen, A., & Marelli, M. (2016). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science*. doi:10.1111/cogs.12392

Hintzman, D. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review, 93*(4), 411–428.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods, 45*(3), 718–730.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review, 23*(16), 1744–1756.

Hollis, G. & Westbury, C. F. (2017). *Identifying dimensions of communicative importance using distributional semantic models*. The High Performance Computing Symposium, 2017. Kingston.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46,* 269–299.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41*(6), 401–410.

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual variability on word learning. *Psychonomic Bulletin & Review, 23*(4), 1214–1220. doi:10.3758/s13423-015-0980-7

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *The Journal of the Acoustical Society of America, 132*(2), EL74–EL80.

Johns, B. T., & Jones, M. N. (2010, August 11–14). *Are random representations accurate approximations of lexical semantics?* Paper presented at the 32nd Meeting of the Cognitive Science Society, Portland, OR.

Johns, B. T., & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 69*(3), 233.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 66*(2), 115.

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

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods, 42*(3), 643–650.

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.

Kuperman, V., Drieghe, D., Keuleers, E., & Brysbaert, M. (2013). How strongly do word reading times and lexical decision times correlate? Combining data from eye movement corpora and megastudies. *The Quarterly Journal of Experimental Psychology, 66*(3), 563–580.

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

Lai, S., Liu, K., Xu, L., & Zhao, J. (2016). How to generate a good word embedding. *IEEE Intelligent Systems, 31*(6), 5–14.

Landauer, T. K. (2002). On the computational basis of learning and cognition: Arguments from LSA. *Psychology of Learning and Motivation, 41,* 43–84.

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.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2177–2185). Cambridge: MIT Press.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203–208.

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

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*(5), 375.

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

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. Retrieved from https://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26,* 3111–3119.

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)* (pp. 746–751). Stroudsburg, PA: Association for Computational Linguistics.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111*(3), 721.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*(2), 327.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory, 2,* 64–99.

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM, 8,* 627–633.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104,* 192–233.

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

Shaoul, C., & Westbury, C. (2010a). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods, 42,* 393–413.

Shaoul, C., & Westbury, C. (2010b) *The Westbury Lab Wikipedia corpus.* Edmonton, AB, Canada: University of Alberta. Retrieved from http://www.psych.ualberta.ca/~westburylab/downloads/westburylab.wikicorp.download.html

Shaoul, C., & Westbury, C. (2013). A reduced redundancy USENET corpus (2005–2011). *University of Alberta, 39*(4), 850–863.

Steyvers, M., & Griffiths, T. L. (2008). Rational analysis as a link between human memory and information retrieval. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 329–349). Oxford: Oxford University Press.

Tainturier, M. J., Tremblay, M., & Lecours, A. (1989). Aging and the word frequency effect: A lexical decision investigation. *Neuropsychologia, 27*(9), 1197–1202.

Wallot, S. (2014). From "cracking the orthographic code" to "playing with language": Toward a usage-based foundation of the reading process. *Frontiers in Psychology, 5,* 891. doi:10.3389/fpsyg.2014.00891

Wallot, S., Hollis, G., & van Rooij, M. (2013). Connected text reading and differences in text reading fluency in adult readers. *PLoS ONE, 8*(8), e71914. doi:10.1371/journal.pone.0071914

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207.