Behav Res (2015) 47:666–684
DOI 10.3758/s13428-015-0614-z

Performance impact of stop lists and morphological decomposition on word–word corpus-based semantic space models

Jeff Keith [1,2] · **Chris Westbury** [1] · **James Goldman** [1]

Published online: 23 June 2015

Abstract Corpus-based semantic space models, which primarily rely on lexical co-occurrence statistics, have proven effective in modeling and predicting human behavior in a number of experimental paradigms that explore semantic memory representation. The most widely studied extant models, however, are strongly influenced by orthographic word frequency (e.g., Shaoul & Westbury, *Behavior Research Methods, 38*, 190–195, 2006). This has the implication that high-frequency closed-class words can potentially bias co-occurrence statistics. Because these closed-class words are purported to carry primarily syntactic, rather than semantic, information, the performance of corpus-based semantic space models may be improved by excluding closed-class words (using stop lists) from co-occurrence statistics, while retaining their syntactic information through other means (e.g., part-of-speech tagging and/or affixes from inflected word forms). Additionally, very little work has been done to explore the effect of employing morphological decomposition on the inflected forms of words in corpora prior to compiling co-occurrence statistics, despite (controversial) evidence that humans perform early morphological decomposition in semantic processing. In this study, we explored the impact of these factors on corpus-based semantic space models. From this study, morphological decomposition appears to significantly improve performance in word–word co-occurrence semantic space models, providing some support for the claim that sublexical information—specifically, word morphology—plays a role in lexical semantic processing. An overall decrease in performance was observed in models employing stop lists (e.g., excluding closed-class words). Furthermore, we found some evidence that weakens the claim that closed-class words supply primarily syntactic information in word–word co-occurrence semantic space models.

Keywords Semantic space model · Vector space model · Semantic memory · Lexical co-occurrence · Stop lists · Closed-class words · Morphological decomposition · Part of speech tagging · Word stemming

> *You shall know a word by the company it keeps*. (Firth, 1957, p. 11)

Human language, and the semantic representation it facilitates, is a complex behavior. To understand language, one needs to know the meaning of words, and retain knowledge regarding the grammatical application of words. The former requirement is addressed by *lexical semantics*, or the study of individual word meanings as constrained by morphology. Here, meaning is defined by context that is likely derived from statistical redundancies in multisensory elements perceived in environment—that is, more than those found in analyzing text alone. Using text alone is not likely to ever provide a comprehensive basis for modeling language comprehension, yet, it has been shown that many aspects of perception and cognition can be understood in isolation by modeling specific capacities as computational problems (Anderson, 1990; Marr, 1982). One such approach in acquiring an understanding of semantic representation involves using simple mechanism(s) operating on large scale. This approach has yielded a rich history of both high level and derived mechanistic memory models for lexical semantic representations. Many of these mechanistic models

✉ Jeff Keith
jkeith1@ualberta.ca

[1] University of Alberta, Edmonton, Alberta, Canada

[2] Department of Psychology, University of Alberta, Edmonton, Alberta T6G 2R3, Canada

can be used in higher-order models of language comprehension (e.g., Kintsch, 1998, 2001).

Semantic space models define a word space in which individual words are represented as points in the space, with their relative locations being defined by their relatedness to the dimensional anchors. These models build upon Osgood's (1952) early multidimensional representation approach (see also Osgood, Suci, & Tannenbaum, 1957). In their early forms, however, they relied on a limited number of human judgments about the semantic nature of words to derive a set of dimensions.

Current semantic space models build lexical semantic representation directly from statistical co-occurrence of words in a corpus of text, storing these representations in a corpus-derived high-dimensional semantic space. The use of statistical co-occurrence enables unsupervised learning from text, minimizing assumptions made in processing and representation. It also provides distributed representations for words, whose meaning is the aggregate distributed pattern of all abstract dimensions, which have no interpretable meaning on their own. In other words, dimensions combine to form an irreducible context. Another virtue of semantic space models is that they are able to reveal latent semantic relationships: words sharing similar context (e.g., two nouns with common features, such as *dog* and *cat*, or *Paris* and *Tokyo*) are regarded as being semantically related, even if they rarely co-occur in the same sentence.

Corpus-based semantic space models

Corpus-based models of semantic representation (also known as *semantic spaces*, *vector spaces*, *word spaces*, or *distributional semantic models*) characterize the meaning of linguistic expressions in terms of lexical distributional properties. These models are commonly unstructured, and capture primarily attributional—and, indirectly, relational—similarity (e.g., can capture standard taxonomic semantic relationships such as hypernymy, synonymy, and co-hyponymy; Turney, 2006).

This section provides a brief functional overview of the major corpus-based semantic space implementations, many of which were built upon earlier work involving forming vector representations of word meaning (Schütze, 1993; Schvaneveldt, 1990). Beyond brief functional descriptions, this overview is intended to highlight the persistence of undesirable orthographic word frequency effects in various implementations of corpus-based semantic space models (about which more will be said later).

Latent semantic analysis (LSA; Landauer & Dumais, 1997) has arguably received the most attention of all the semantic space model implementations. LSA starts by computing a *word* × *document* frequency matrix from a large corpus of text, resulting in a very sparse matrix. The row entries (word vectors) capture the frequency of each word in a particular document, and are normalized using an entropy function.[1] Next, the dimensionality of the sparse matrix is reduced using singular value decomposition (SVD; a form of factor/principle component analysis), which brings out latent semantic relationships between words, even if they have never co-occurred in the same document. The basic premise behind LSA is that the aggregate contexts in which a word does and does not appear provide a set of constraints to induce the word's meaning (Landauer, Foltz, & Laham, 1998). LSA has been criticized for being prone to the influence of strong orthographic word frequency effects (despite its entropy based normalization) on vectors and vector distances (Rohde, Gonnerman, & Plaut, 2006), and as being a "bag of words" model, ignoring statistical information inherent in word transitions within documents (Perfetti, 1998).

The *Hyperspace Analog to Language* (HAL; Burgess, 1998; Lund & Burgess, 1996) implementation uses word co-occurrence in a corpus to build an abstract data representation (i.e., a vector space captured by a *word* × *word* matrix) that contains contextual information for every word in a specified dictionary. HAL uses N dimensions for this vector space (i.e., yielding an $N \times N$ matrix), where N equals the number of words in the dictionary used while processing the corpus. Each vector created represents a word from this dictionary. Each column entry in a word's vector is a count of the number of times it co-occurred with another word in the corpus, weighted by a factor that specifies how distant the two words were on each co-occurrence. In the original HAL implementation, words were considered to have co-occurred if they appeared within ten words of each other, in either direction (i.e., a window size of 10F/10B). Lexical semantic memories are built by reading words one window at a time, counting co-occurrences, and then sliding the window forward one word. Vectors with the lowest row variances (i.e., sparse rows) are eliminated. Minkowskian distance metrics (e.g., Euclidean or *city-block*) are used to calculate the distance between any two word vectors in the space, and because these metrics are sensitive to vector magnitudes, vectors are algebraically normalized to unit vectors. If the words have similar values in the same dimensions, they will be closer together in the space, meaning they share similar contexts and are presumably semantically related. The word vectors closest to a given word are considered its neighbors.

[1] Entropy, in information theory, is a measure of the information content of a token—such as a word—in a given context; in the LSA model's use of entropy, the more evenly distributed a word is across documents (i.e., the more frequent it is), the lower its weighting in the model, following the intuition that frequent words are less informative.

The *High Dimensional Explorer* implementation (HiDEx; Shaoul & Westbury, 2010; 2012), an extension of HAL, was developed to address three major shortcomings of HAL. First, despite its vector normalization, HAL is prone to the influence of strong orthographic word frequency effects on vectors and vector distances (Shaoul & Westbury, 2006), and HiDEx uses improved vector normalization schemes to counter these effects. Second, the majority of the parameters used in HAL (e.g., window size, co-occurrence distance weighting, distance/similarity metrics, etc.) were set without any explicit a priori justification; HiDEx allows for manipulation of these parameters. Finally, a configurable neighborhood size and membership threshold is implemented that better accounts for variance in the number and average "closeness" of neighbors between different words, providing more meaningful neighborhood density measurements.

Windsor Improved Norms of Distance and Similarity of Representations of Semantics (WINDSORS; Durda & Buchanan, 2008) is another extension of the HAL implementation. WINDSORS's primary aim is to eliminate orthographic word frequency effects and uses two mathematical techniques to do so: log-relative frequency ratios (Damerau, 1993) to address high-frequency effects and a simplified Good–Turing correction[2] (Good, 1953) to address low-frequency effects.

The *Correlated Occurrence Analogue to Lexical Semantic* implementation (COALS; Rohde et al., 2006) is also based on HAL. COALS achieves significantly better performance over HAL through improvements in vector normalization, again, to address orthographic frequency effects. Specifically, it converts raw co-occurrence counts to Pearson correlations between each target word and the other words in the lexicon. These correlation values tend to be very small (i.e., $1 >> r > 0$), and so are square-rooted. Any negative correlations, which are regarded by the model as being largely uninformative, are set to a value of zero. This increases the sparseness of the co-occurrence matrix, thereby optimizing the subsequent dimensionality reduction performed through SVD.

Bound Encoding of the Aggregate Language Environment (Jones & Mewhort, 2007) analyzes a single sentence at a time, recording each word's context (i.e., co-occurrence) and the word order information. After processing an entire corpus, context and order information are combined into single word vectors using a circular convolution function,[3] and orthographic word frequency is controlled for using an entropy function (similar to LSA) for vector normalization.

Limitations of corpus-based semantic space models

Closed-class words—that is, function words such as determiners (e.g., *the*, *a*, etc.) and common prepositions (e.g., *to*, *for*, etc.)—have much higher orthographic frequencies than the rest of the words in a language. It is common practice in the information retrieval literature to use stop lists (i.e., lists of high-frequency words, such as closed-class words, that are excluded by the model), as the highest frequency words are regarded as semantically uninformative as context dimensions (Manning & Schütze, 1999; Rapp, 2003; Smith & Humphreys, 2006). Moreover, discarding them greatly reduces both the corpus size (often by up to 50%) and, to a lesser degree, the computing requirements for the co-occurrence statistics. Bullinaria and Levy (2007), however, found that doing so reduces—or at the very least offers no significant improvement in Bullinaria and Levy (2012)—performance of word–word co-occurrence models.

The COALS implementation excludes closed-class words and it has been shown doing so leads to better performance over HAL, and, importantly, that adding closed-class words back into COALS reduced its performance (Rohde et al., 2006). BEAGLE uses stop lists of closed-class words for context information, but not for word order information. LSA, a word-document model, generally includes closed-class words in the initial co-occurrence data (i.e., before dimension reduction), but when LSA is used to compare word strings shorter than normal text paragraphs (e.g., short sentences) zero weighting of function words (i.e., excluding closed-class words) is often pragmatically useful (Landauer & Dumais,

[2] The Good–Turing methods provide estimates of the total probability of unseen events. Where a target word does not co-occur with a given word in a corpus, and therefore has a co-occurrence value of zero (i.e., an *unseen event*), the log-relative frequency ratios would be undefined; the Good-Turing method is used to estimate non-zero values for such *unseen events*, ensuring the log-relative frequency ratios are well-defined for all entries in the co-occurrence space.

[3] Also known as a *holographic* model because it is based on the same mathematical principles as light holography. When BEAGLE combines context and word order information (stored in individual vectors), it calculates the outer-product between vectors, which results in the *binding problem*: the resultant outer-product vector has more dimensions than its parents (i.e., $[kn - k - 1]$ dimensions, where k = number of vectors combined and n = number of dimensions in each parent vector); meaning, many semantic context dimensions would be added to the model, solely as an artifact of a mathematical manipulation. The *circular* convolution function in BEAGLES calculates the outer-product and returns a vector of the *same* length as its parents, while minimizing information loss by algorithmically convoluting—or reflecting—the information from the removed outer-product dimensions back into their retained neighbours.

2008). All other implementations discussed typically include closed-class words in their co-occurrence statistics.

In the corpus-based semantic space literature, it has been suggested that closed-class words, which are typically found in close proximity to target words (Shaoul & Westbury, 2010), are more syntactically related, rather than semantically related, to target words (Rohde et al., 2006). Unstructured co-occurrence models of lexical semantics generally ignore syntactical information in meaning creation, yet when human subjects perform semantic judgment tasks, they rely, in part, on syntactic properties such as word class (Rohde et al., 2006).

It may prove useful to use stop lists to exclude closed-class words while including part-of-speech (i.e., syntactic) information about words in semantic space models. This assertion is made, given (1) the inconsistent results with closed-class exclusions reported between models, (2) the confounding influence of high orthographic frequency on co-occurrence statistics, and (3) closed-class words' purported syntactic role.

Another significant limitation of corpus-based semantic space literature is that very little work has been done to investigate the performance impact of carrying out a full morphological decomposition on target words prior to collecting co-occurrence statistics. In current models, each inflected form of a word (e.g., *happy*, *unhappy*, *happiness*, *happier*, etc.) is represented as an individual vector, making the implicit, and somewhat non-intuitive, assumption that each inflected form has distinct semantic representation in human memory. The plausibility of this assumption is called into question by evidence for morphological decomposition occurring early in the visual processing of words (e.g., Solomyak & Marantz, 2010).

One key study used morphological decomposition with corpus-based semantic space models; this study reported that morphological decomposition afforded no significant improvement to performance of word–word co-occurrence models (i.e., HAL-based implementation; Bullinaria & Levy, 2012). However, in this study only partial morphological decomposition was accomplished and evaluated: through simple word stemming, and by using a standard lemmatized version of their corpus. Both stemming and lemmatization reduce inflected word forms to a common base form—not necessarily the inflected forms' monomorphemic stem. Stemming was carried out using Porter's (1980) algorithm, which employs a basic, heuristic approach that removes many word suffixes and derivational affixes. Lemmatization usually refers to decomposing words "properly" through the use of a predefined vocabulary and morphological analysis of words, but it is difficult to perform accurately for very large corpora because of its dependency on word context. Additionally, no analysis was carried out in Bullinaria and Levy's study to determine whether retaining stripped affixes as part of context had any effect on resulting performance.

There is some support for morphological decomposition having an impact on co-occurrence model performance. Jing and Tzoukermann (1999) used externally provided morphological information and showed it improved calculations of semantic relatedness between two words (i.e., by computing the distance between their vectors using an implementation based on second-order, word–word lexical co-occurrence statistics). Harman (1991) found that stemming provided no performance improvement, regardless of the stemming algorithm used. Krovetz (1993), on the other hand, showed that stemming improved performance in various tasks by up to 45.3%. Hull (1996) similarly concluded that stemming is almost always beneficial, but, disagreeing with Krovetz, claimed the average improvement due to stemming is only 1% to 3%.

Moving away from word–word co-occurrence models, the use of morphological decomposition has been explored elsewhere. Using the same word-stemming approach taken by Bullinaria and Levy (2012), Landauer, McNamara, Dennis, and Kintsch (2007)—using a standard LSA implementation—found that stemming offered no improvement; indeed, Landauer and Dumais (2008) claim that "stemming often confabulates meanings" (p. 4356) in word vectors.

It remains an outstanding question as to whether a *full* morphological decomposition (i.e., stripping both prefixes and suffixes as accurately and completely as possible) of the corpus text will have a significant impact on the predictive capabilities of word–word co-occurrence statistics in semantic space models.

Current study

The present study explored the individual and interactive effects of using stop lists (i.e., to exclude closed-class words), limited syntactic information (via part-of-speech [POS] tagging and/or retaining stripped affixes as context dimensions), and morphological decomposition on word–word co-occurrence semantic space models. These factors have been only partially addressed—or, at times, completely ignored—in the corpus-based semantic space literature. Various combinations of these factors were used to build corpora from which co-occurrence statistics were computed, and compared with each other by looking for significant differences in performance on a set of semantic tasks.

A number of potential outcomes were foreseen. For example, it was expected that through using POS tags and/or affixes for context, the syntactic information presumably captured by closed-class words would be retained, while excluding closed-class words would eliminate their orthographic frequency effect, possibly leading to an overall improvement in performance. In fact, some of the earliest work on co-occurrence statistics from large corpora (e.g., Finch & Chater, 1992) was actually focused on defining syntactic categories, rather than semantics, suggesting that the word vectors already contain a combined representation of both semantics and syntax.

In using a morphologically decomposed corpus, all contextual information for each inflected form of a word is combined into a single, monomorphemic lexical stem vector. As suggested by Landauer and Dumais (2008), this compression of information may introduce excessive noise into the context, thereby reducing semantic content. Implicit in such a view, however, is the assumption that compressing—or "confabulating"—co-occurrence statistics into a combined representation would necessarily result in a loss of information. The authors are not aware of any mathematical proofs and/or empirical support for this implied information loss resulting from combining dimensions. In the present study, it was also considered that a morphological decomposition might yield a richer semantic representation of words. Rapp (2003) developed a relatively simple algorithmic machine translation approach to inducing context-dependent word sense (e.g., disambiguating homographs) from co-occurrence statistics. In doing so, Rapp demonstrated that word vectors in such models contain an aggregate representation of the underlying semantics, which implies content rich vectors (i.e., those collecting co-occurrence statistics for a word used in multiple contexts an senses) can provide better representations of a given word's lexical semantic content and provide access to machine-driven approaches to understanding language.

We used HiDEx to test the performance impact of stop lists (i.e., to exclude closed-class words), POS tagging, and morphological decomposition on word–word corpus-based co-occurrence semantic space models.

Method

Factor combinations explored

This study used a 2 × 2 × 3 factorial design (see Fig. 1), with the following factors and levels: (1) POS Tagging (with or without), (2) Stop List (used or not used), and (3) Morphological Decomposition (used [with affixes included in lexicon], used [without affixes included in lexicon], and not used). Morphological decomposition included both prefixes and suffixes, and where affixes were included in the lexicon, they were treated as individual words (i.e., vectors and context dimensions) in the co-occurrence space.

Base corpus

The ukWaC corpus (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009) was used in this study. It contains close to two billion words and was built from web content derived from crawling the UK Internet domain. This corpus was chosen because it has a POS-tagged version available. It was also chosen for its size; it is large enough to provide more reliable statistics and better performance than smaller, higher content quality corpora (Bullinaria, 2008; Bullinaria & Levy, 2007) while still being small enough to make its use in 12 different models computationally efficient.

Because the ukWaC corpus is derived from Web content in the UK Internet domain, spelling of words in our corpora and target lists were both subjected to the same UK-to-US word form replacement process. A comprehensive list of 2,282 UK-to-US spelling variants (Tysto, 2012; Wikipedia, 2014) was used to make these replacements. Although the differences between British and American English go beyond spelling (e.g., unique word usages and differing patterns of word co-occurrences), these differences were considered minor for the purposes of this study, and as being outweighed by the other advantages offered by this corpus.

Morphological decomposition

Full morphological decomposition (i.e., of both prefixes and suffixes) was carried out by the parsing sub-functions from the freely available PC-KIMMO (SIL International, 1997) application (see also, Koskenniemi, 1984; Sproat, 1991). These sub-functions are called from within PrepCorpus, a custom-built application, which incorporates new functions to optimize performance and accuracy. The base accuracy of PC-KIMMO is estimated[4] at approximately 95% (i.e., it performs correct morphological decomposition and stem replacement on approximately 95% of inflected words), with 1% error parses (i.e., words that were parsed, but parsed incorrectly) and 4% missed parses (i.e., inflected words that PC-KIMMO failed to parse). Using PrepCorpus to handle both systematic and idiosyncratic errors and misses, we achieved just over 99% accuracy[5] in our morphological decompositions. Lastly, where irregular root forms were encountered, lexical stem replacements were used in the morphological decomposition (e.g., *spies* replaced with *spy* +*s*), and affixes, which are output as stand-alone "words," were identified by concatenating them with a "+" symbol (e.g., *unknowable* is output as *un*+ *know* +*able*) in order to disambiguate them from certain word stems (e.g., to distinguish between instances of the word *able* and the suffix +*able*). The base ukWaC corpus has just over 1.99 billion words; morphologically decomposing, while retaining affixes as independent tokens, resulted in an increased corpus token count of 2.41 billion words.

[4] These estimates were derived from manually reviewing random blocks of text (approximately 30,000 words in total) in a parsed sample corpus.

[5] Calculated by taking six random samples (50,000–70,000 words each) of text from the morphologically decomposed corpora and tabulating the number of errors, misses and parses in each.

		Morphologically Decomposed?		
		YES with affixes	YES without affixes	NO
With POS-tagging	Stop-list used (i.e. closed class words excluded)	1	5	9
	Stop-list not used (i.e. closed class words included)	2	6	10
Without POS-Tagging	Stop-list used (i.e. closed class words excluded)	3	7	11
	Stop-list not used (i.e. closed class words included)	4	8	12 baseline

Fig. 1 Depiction of the 2 × 2 × 3 factorial combinations explored in this study

For reasons outlined in this article's Discussion section, compound words (e.g., *sunshine*, *grandmother*, *scarecrow*, etc.) were not decomposed in this study.

POS tagging

The POS tagging of the ukWaC corpus was performed using TreeTagger (Baroni et al., 2009), and PrepCorpus, enabling a POS tag to be added to the lexical stem of parsed inflected word forms (e.g., *unhappiness* tagged as an adjective in the original ukWaC corpus, is output as *unhappiness*_J in those corpora without morphological decomposition, and *un*+ *happy*_J +*ness* in POS-tagged and morphologically decomposed corpora). With POS tags, when the same word is used in different syntactic roles (e.g., homographs, such as *saw*, which can be used as a verb or noun, depending on the context), each distinct syntactic use was represented as a different vector in the model (i.e., *saw*_N and *saw*_V). We used four distinct POS tags: nouns (_N), verbs (_V), adverbs (_A), and adjectives (_J).

Stop lists and closed-class word exclusions

The stop list used in this study was comprised of 323 words. It was preliminarily constructed by taking all single words (i.e., nonphrase) from the list of closed-class words compiled by, and freely available from, Sequence Publishing (2014), which includes auxiliary verbs, conjunctions, determiners, prepositions, pronouns, and quantifiers (total of 216 words). To this list, the most frequent words in the raw ukWaC corpus were added (total of 107 words), which either belonged to a closed-class (e.g., inflected forms not included on the Sequence Publishing lists) or that were nonsensical tokens (e.g., occurrences of single letters such as *n*, *p*, *x*, etc.). Removing all word tokens included on the stop list reduced the corpus token count to 1.04 billion words.

Co-occurrence model and parameterization

HiDEx was used to process all 12 corpora reflecting the factor combinations in this study (see Table 1 for HiDEx parameterization. Bullinaria and Levy (2007) have shown that using positive pointwise mutual information for vector normalization, and a cosine similarity metric, optimizes performance of word–word corpus-based semantic space models. These authors also found that using an unweighted window size of one word (in either direction) was optimal. Given that one of the aims of this study was to compare the performance impact of retaining stripped affixes as context dimensions, a larger window size was deemed appropriate (i.e., to handle inflected word forms with more than one prefix, like unpremeditated [*un*+ *pre*+ meditate +*ed*], or more than one suffix, like computerization [compute +*er* +*ize* +*ation*], while still capturing co-occurrences with adjacent words, which could also be surrounded by their own stripped affixes). Bullinaria and Levy (2007, 2012) found that window sizes of up to three words in either direction still performed very well (i.e., major drops in performance were not noted in most of the tasks until reaching window sizes of four or five), and Shaoul and Westbury (2010) found that a window size of five words in either direction (using a linear ramp weighting scheme) led to optimal performance in a lexical decision task (also used in this study). We chose a window size of four words in either direction to accommodate affix retention, while minimizing the trade off in performance. With that all said, because we considered comparisons of performance between our various factor combinations, achieving absolutely optimal performance was less important than selecting parameterization settings that could accommodate each factor combination.[6]

Lexicon construction

HiDEx takes a user-supplied lexicon as input; a word vector is created for each word both listed in the lexicon and found in the corpus. The lexicon used in each model differed slightly depending on whether stop lists, morphological decomposition, affix exclusions, and/or POS tagging was used in a given model. The need for custom lexicons for each model can be simply illustrated by considering the lexicon entry *deregulation*; in some corpora this word will be represented in its full form, but in others, it would appear as *de*+ *regulate* + *tion*, *regulate*, *deregulation*_N, *de*+ *regulate*_N +*tion*, or

[6] For more elaborate investigations and discussions on model parameterization and metrics please refer to Bullinaria and Levy (2007, 2012), Kiela and Clark (2014), and Lapesa and Evert (2013).

Table 1 HiDEx settings used for computing co-occurrence statistics

Parameter	Configuration Used
Corpus	ukWaC, with various manipulations
Normalization	Positive pointwise mutual information
Vector similarity metric	Cosine
Weighting scheme	Linear ramp
Window size	Four words in either direction
Context size (*N*)	15,000 (×2, forward/backward)

*regulate*_N. The average lexicon size was 62,573 words—with a maximum of 64,806 words in the base corpus, and a minimum of 60,772 words in the morphologically decomposed corpus with closed-class and affix exclusions.

Tasks used for model comparisons

Each of the 12 sets of co-occurrence statistics resulting from the factor combinations were used to complete three separate semantic tasks:

- ***Distance comparison (DC)*** For this task, we used 200 pairs of semantically related words (e.g., *thunder* and *lightning*, *black* and *white*, *brother* and *sister*). The distance between each target word (i.e., first word in each pair) and the second word in the pair was computed. Then the distances between the target word and each of ten other randomly selected control words from the other 199 pairs was computed. Performance was measured by counting the number of times a given semantic pair was closer in the semantic space than the first word in that pair and each of the ten control words. This task is similar to the often-studied *Test of English as a Foreign Language*[7] (TOEFL) test originally used by Landauer and Dumais (1997), but makes use of more frequent, better-distributed words than TOEFL (Bullinaria & Levy, 2007). It tested each corpus-based semantic space model's ability to perform semantic similarity judgments.
- ***Semantic categorization (SC)*** For this task, we used ten words from each of 53 semantic categories (e.g., precious stones, units of time, familial relationships, vegetables) based on Battig and Montague's (1969) category norms. A category center was calculated for each category by taking the mean of the vectors corresponding to the last nine words in each category, and performance was evaluated by counting the number of times the first word in a given category is closer to its category center than the category centers of the other 52 categories. This task tested each corpus-based semantic space model's ability to represent known semantic categories (Patel, Bullinaria, & Levy, 1997).
- ***Lexical decision (LD)*** This task was used to model human behavior in an LD task. Rather than using data from a limited number of subjects, the aggregate data from the English Lexicon Project was used (Balota et al., 2007), which included LD reaction time data for over 32,000 words. Along with the log-transformed orthographic frequency, the ARC and INV-NCOUNT neighborhood measures (Shaoul & Westbury, 2012) from HiDEx were used as LD reaction time predictors in linear regression models. Performance was evaluated by comparing each model's change in variance accounted for in the data, relative to the base model's performance (i.e., ΔR^2).

Tasks similar, but not identical, to the DC and SC tasks described above were used by Bullinaria and Levy (2007, 2012) in their studies exploring the effects of stop lists and word stemming on corpus-based semantic space models. The same word sets were used for those tasks in this study (Bullinaria, 2013). The DC and SC tasks are used here in order to enable a very rough comparison between approaches to morphological decomposition. One performance evaluation used for these tasks by Bullinaria and Levy (2007, 2012) was a simple comparisons of each task's count proportions as percentages. This measure was also used here. It is not, however, an ideal measure because a very small amount of variance was expected between model performance percentages (i.e., there was a ceiling effect), and the measure is not amenable to rigorous significance testing. Instead, to test for statistical significance, Bullinaria and Levy split their corpus into disjoint subsets and performed *t* tests between manipulations, comparing the mean performance of each corpus subset for each corpus manipulation. However, in the earlier study, Bullinaria and Levy (2007) showed that using smaller corpora resulted in inferior performance. As such, in the present study, each task's count proportions were also modeled using beta-distribution linear regressions (Ferrari & Cribari-Neto, 2004), and models were compared using Akaike information criterion (AIC) relative likelihoods.

Relative to the DC and SC tasks, LD is a more difficult task for co-occurrence models—they typically account for slightly less than a third of the variance in LD reaction times (Buchanan, Westbury, & Burgess, 2001; Shaoul & Westbury, 2010). As such, the LD task was expected to provide a better scale—that is, a greater range of performance—to reveal differences between the 12 factor combinations.

[7] The TOEFL is considered by some to have become a standard benchmark task for semantic space models, but, as it is copyrighted material, it is not freely available and the authors of this study were unable to either obtain a copy of it, nor permission to use it.

Beta-distributed linear regression models

In both the DC and SC tasks, a more statistically rigorous approach to comparing factor combinations was needed—one that made use of all of the data available. Treating the performance percentages as count data, a generalized linear model using a Poisson distribution is typically prescribed. In the present study, however, DC and SC response data were discrete (noncontinuous), heavily skewed, heteroskedastic, failed a chi-squared goodness of fit test for Poisson distribution (i.e., did not follow Poisson distribution), and had an unequal mean and variance. Because they specifically address the DC and SC response data properties outlined, regression models were built on the basis of beta distributions (Ferrari & Cribari-Neto, 2004).

Target list construction

Custom target lists for each of the three tasks were created for each of the 12 factor combinations compared, converting targets into forms appropriate for each (i.e., morphologically decomposed and/or POS-tagged). Additional refinement of the custom LD targets (and lexical decision reaction time [LDRT] data) was required for morphologically decomposed models, in order to account for many cases in which multiple inflected forms of the same lexical stem were present in the original ELP LDRT data set. In those cases in which a lexical stem was present on the original target list (e.g., "abandon"), all other records for inflected forms of that stem were disregarded (e.g., "abandoned," "abandoning," "abandonment"), both on the custom target list and in the LDRT data used for the model.

This process yielded an average of 32,124 custom targets for nondecomposed factor combinations (e.g., base: DC custom/original targets = 199/200 pairs; SC custom/original targets = 529/530 words; LD custom/original targets = 32,290/32,681 words), and 15,700 targets for morphologically decomposed factor combinations (e.g., morphologically decomposed [only]: DC custom/original targets = 199/200 pairs; SC custom/original targets = 528/530 words; LD custom/original targets = 15,801/15,975 words).

Results

Factor combination naming convention

The results reported here all make use of a shorthand abbreviation system—shown in Table 2—to identify each of the 12 factor combinations.

DC and SC performance percentages

Performance percentages are reported here for the sake of comparison to Bullinaria and Levy's (2012) results, though the tasks are not identical. These are poor measures, since they are clustered at the high end of the scale's range (i.e., pronounced ceiling effect), and given that very little performance variation was observed between models. In the SC task, only 0.00002% variance was observed between all factor combinations' performance relative to the base; similarly, in the DC task the observed variance was 0.00067%. Despite this shortcoming, performance differences between factor combinations were noted, and the performance rankings in DC and SC tasks correlated reliably with each other (Spearman's rank-ordered correlation, $r = .74$, $p = .006$).

Performance percentage results for the DC task are summarized in Fig. 2. On the basis of these data, the **morph** (morphological decomposition only), **pos** (POS tagging only), and **pos.morph** (both POS-tagged and decomposed) factor combinations all performed better than baseline.

Performance percentage results for the SC task are summarized in Fig. 3. On the basis of these data, the **pos.morph.ax** (POS-tagged, decomposed, and with affixes removed) and **pos.morph** (both POS-tagged and decomposed) factor combinations performed better than baseline.

In the DC task, this study's morphologically decomposed factor combinations performed worse (80%–86%) than Bullinaria and Levy's (2012) stemmed and lemmatized models (92%–93%; p. 896). On the other hand, decomposed factor combinations performed better on the SC task in the present study (95%–96%) than in Bullinaria and Levy's[8] (80%–82%; p. 896). These comparisons were made with Bullinaria and Levy's (2012) models using 15,000 context dimensions, the same number of context dimensions used in this study.

DC beta-model performance

In order to facilitate the DC task's earlier discussed beta-distributed regression analysis, the response variable (i.e., the number of comparisons in which the target was closer to its own pair word than to ten random control words) was transformed to proportions. As Smithson and Verkuilen (2006) recommended, a betareg correction was then applied in order to eliminate response values of 0 and 1 (i.e., replaces

[8] It should be noted, however, that the SC task implemented by Bullinaria and Levy (2012) differed from ours. We used one categorization test for each of the 53 categories, whereas Bullinaria and Levy used one categorization test for each of the 530 words; this could account for the large increase in baseline performance noted in the present study.

Table 2 Factor combination abbreviations used

		No Morphological Decomposition	**morph**: Morphologically Decomposed	
			Affixes *Included* in Corpus	**ax**: Affixes *Excluded* From Corpus
Closed-class words *included* (i.e., no stop list)	**pos**: *with* POS tagging	**pos**	**pos.morph**	**pos.morph.ax**
	without POS tagging	***base***	**morph**	**morph.ax**
cx: Closed-class words *excluded* (i.e., uses stop list)	**pos**: *with* POS tagging	**pos.cx**	**pos.morph.cx**	**pos.morph.cx.ax**
	without POS tagging	**cx**	**morph.cx**	**morph.cx.ax**

with approximations—e.g., 0 becomes .000232, and 1 becomes .9934). Models[9] were built using the betareg function from the *betareg* package (Cribari-Neto & Zeileis, 2010) in the statistical computing environment R (R Development Core Team, 2013).

The AIC value for each betareg model was calculated and compared via relative likelihoods, calculated as

$$\text{R.L.}_{basevs.model} = e^{(\text{AICmodel}-\text{AICbase})/2}.$$

The results—summarized in Fig. 4—show that the **pos.morph** (both POS-tagged and decomposed), **morph** (morphological decomposition only), and **pos** (POS tagging only) factor combinations all performed significantly better than baseline.

SC beta-model performance

In order to facilitate the SC task's earlier discussed beta-distributed regression analysis, the response variable (i.e., number of comparisons in which the target is closer to its own semantic category center than to the other 52 semantically unrelated category centers) was transformed to proportions and modeled[10] as outlined earlier.

AIC values for each betareg model were calculated and compared using relative likelihoods. The results are summarized in Fig. 5 and show that the **morph** (morphological decomposition only) and **morph.ax** (morphological decomposition with affixes removed from the corpus) factor combinations both performed significantly better than baseline.

LDRT performance

Targets and their LDRTs from the English Lexicon Project were used to build linear regressions[11] for each model. Because each factor combination had a different number of observations (see the Target List Construction in the Method section), interfactor combination comparison was made difficult, owing to most linear model comparison techniques requiring balanced data. As such, the comparison metric used here was the change in the variance accounted for in a factor combination's regression model, relative to the baseline—that is, $\Delta R^2 = R^2_{\text{model}} - R^2_{\text{base}}$.

The results based on using all available LDRT targets/data are shown in Fig. 6, where it can be seen, somewhat surprisingly, that all models performed worse than baseline, for which R^2 = .31. The **morph** (decomposed only) and **cx** (closed-class word exclusions only) factor combinations performed closest to baseline.

Very high and very low orthographic frequencies (OFREQs) are known to exert a strong influence on lexical access, though the precise nature of this effect is not clear (e.g., Andrews, 1992; Grainger, 1990; McCann, Besner, & Davelaar, 1988; Scarborough, Cortese, & Scarborough, 1977). Given this, only those data associated with target words falling within certain OFREQ ranges were considered.

[9] The betareg formula used predicted corrected.CompPortions by using tarlnOFREQ, pairlnOFREQ, and avgCompslnOFREQ, where tarlnOFREQ is the log-transformed orthographic frequency (OFREQ) of the target, pairlnOFREQ is the logged OFREQ of the semantic pair word, and avgCompslnOFREQ is the average logged frequency of all control words compared against the semantic pair. All predictors entered reliably into all regressions.

[10] The betareg formula used predicted corrected.CompPortions using groupCOS, tarlnOFREQ, and grpAveragelnOFREQ, where groupCOS is the cosine distance between the target and its category center, tarlnOFREQ is the log-transformed orthographic frequency (OFREQ) of the target, and grpAveragelnOFREQ is the average of the logged OFREQ of the all nontarget members of the semantic category. All predictors entered reliably into all regressions.

[11] lm-formula = INV.LDRT ~ lnOFREQ + ARC + lnNCount; where INV.LDRT = the inverse lexical decision reaction time, lnOFREQ = logged orthographic frequency of the target word, ARC = average radius of co-occurrence of target, and lnNCount = logged neighbour count of the target word. All predictors entered reliably into all regressions.

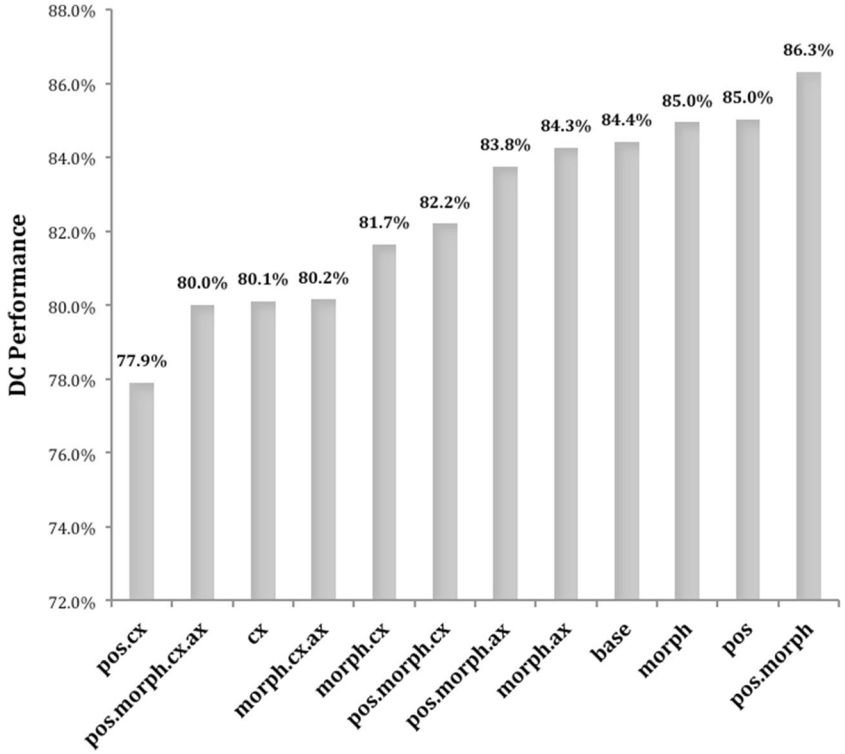

Fig. 2 Rank-ordered performance percentages for each factor combination in the distance comparison (DC) task

Table 3 shows the distribution of the available data across OFREQ bins; the moderate-OFREQ data comprised 58% of the total data available for analysis.

As is shown in Fig. 7, targets with moderate OFREQs accounted for, on average, twice the variance in LDRTs as either the low- or the high-OFREQ data. In further LDRT analyses, only those LD targets with moderate OFREQs were considered.

On the basis of the comparison of these moderate-OFREQ regressions, the **morph.cx** (morphological decomposition with closed-class words excluded), **morph** (morphological decomposition only), and **pos.morph.cx** (pos-tagged and morphologically decomposed with closed-class words excluded) factor combinations each performed better than baseline. The **cx** (closed-class word exclusions only) factor combination's performance was equal to baseline. These results are summarized in Fig. 8.

Aggregate model performance

The performance of factor combinations in each task was ranked by assigning scores to each factor combination based on its performance in that task; that is, factor combinations were assigned a score between 1 and 11 based on their performance relative to baseline, with the best-performing combination being scored 11, the next best scored 10, and so on. The performance rankings of only the most reliable tasks and measures are shown in Fig. 9; that is, we excluded, as unsuitable for rigorous comparisons, the DC and SC task performance percentage measures, as well as the LD task using all available LDRT data (i.e., without target OFREQ filtering), for the reasons outlined earlier. The four top-performing factor combinations are identified as those using morphological decomposition (**morph** and **morph.ax**; with [best] or without [2nd] stripped affixes in the corpus, respectively), POS tagging and morphological decomposition (**pos.morph**; 3rd), and POS tagging, decomposition, and closed-class exclusions (**pos.morph.cx**; 4th). However, only the factor combination using morphological decomposition (**morph**; retaining affixes in the corpus) outperformed baseline in all of the most reliable tasks and measures.

Discussion

In the present study, we explored the performance impact of closed-class word exclusions (using a stop list) and full morphological decomposition on word–word corpus-based co-occurrence semantic space models. We did so by exploring the individual and interactive performance effects of using stop lists, limited syntactic information (via POS tagging and/or affix retention as context dimensions), and morphological decomposition—factors only partially addressed, or completely ignored, in the literature. Various combinations of these factors were used to build corpora, from which co-occurrence statistics are computed and used to compare performance on

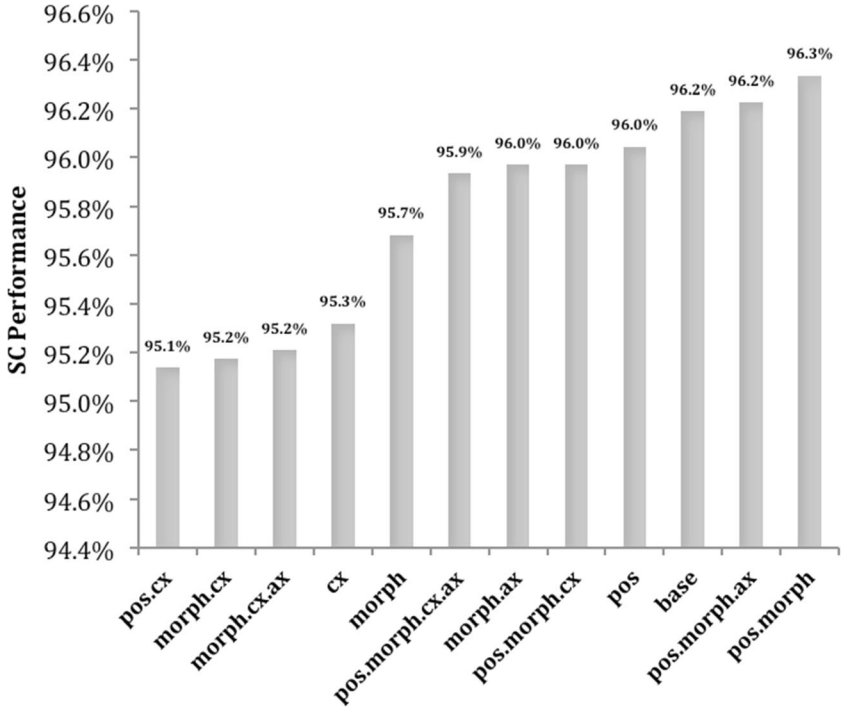

Fig. 3 Rank-ordered performance percentages for all factor combinations in semantic categorization (SC) task

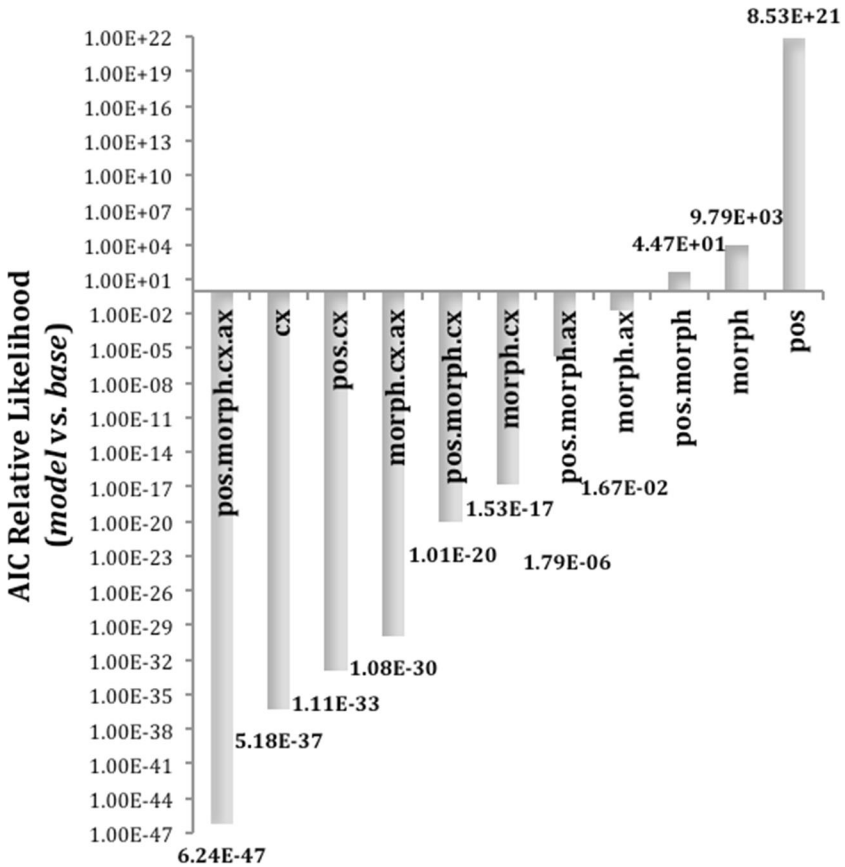

Fig. 4 Akaike information criterion (AIC) relative likelihood comparisons between each distance comparison task beta-regression model and the baseline

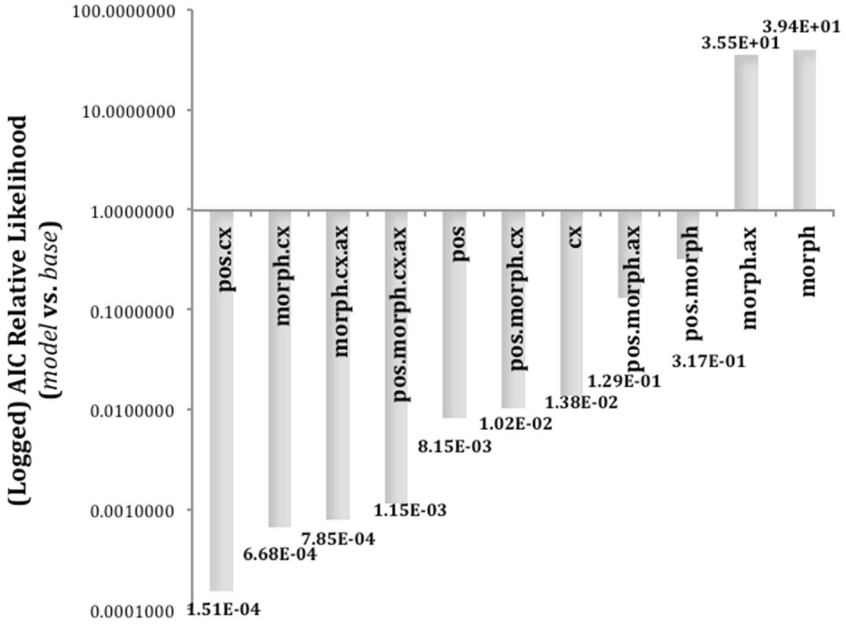

Fig. 5 Akaike information criterion (AIC) relative likelihood comparisons between each sematic categorization task beta-regression model and the base model (AIC relative likelihood values are logged on the *y*-axis for the sake of display convenience, and the data point labels for each model represent the nonlogged values)

semantic tasks—distance comparison, semantic categorization, and lexical decision.

Performance impact of closed-class word exclusions (using stop lists)

We initially speculated that, through using POS tags and/or affixes for context, the syntactic information presumably captured by closed-class words (Rohde et al., 2006) would be retained, and excluding closed-class words would eliminate their contribution to any orthographic frequency effects, possibly leading to an overall improvement in performance.

Using stop lists to exclude closed-class words did not improve performance in the DC and SC tasks; rather, in most cases performance was much worse than baseline. These findings are consistent with Bullinaria and Levy (2012).

Of the four factor combinations that performed equal to, or better than, baseline in the LD task, three employed closed-class exclusions. The best factor combination in that task used POS tagging, morphological decomposition, and closed-class exclusions (which was also the fourth best overall performer). Note, however, that no other factor combination featuring closed-class exclusions performed better than baseline in any task or measure.

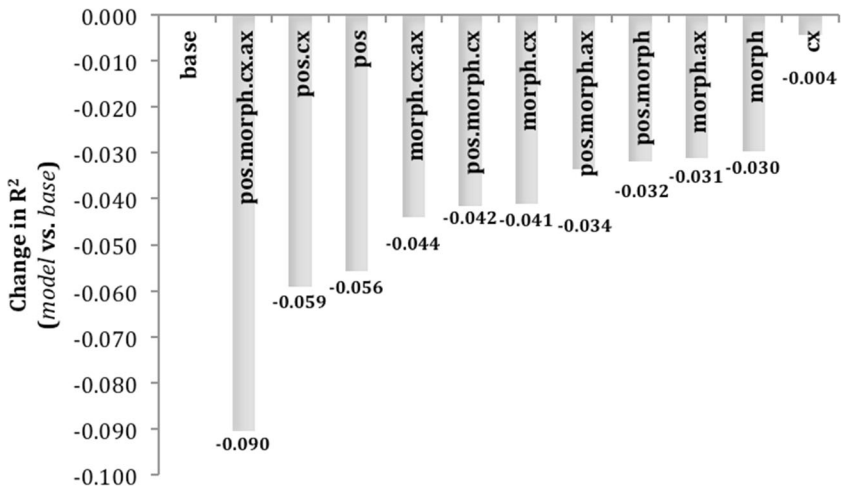

Fig. 6 Change in variance accounted for (ΔR^2, where $\Delta R^2 = R^2_{\text{MODEL}} - R^2_{\text{BASE}}$) between the *full* (i.e., including all data available to each model) lexical decision task regression models and baseline

Table 3 Numbers of words in aggregate lexical decision data in each orthographic frequency bin

Bin[a]	Number of Words	%
1	79,066	32%
2	141,335	58%
3	23,207	10%
Total	243,608	

[a] Bin 1: *OFREQ* < 1; Bin 2: 1 ≤ *OFREQ* < 50; Bin 3: *OFREQ* ≥ 50; occurrences per million words of text

Some interesting findings emerged when considering the interaction between closed-class exclusions and either POS tagging or affix inclusions (the latter in decomposed corpora only). In all tasks and measures, direct comparisons between closed-class exclusive factor combinations varying only in affix inclusion (i.e., **morph.cx** vs. **morph.cx.ax**; **pos.morph.cx** vs. **pos.morph.cx.ax**; see Table 2), factor combinations with affixes performed better than those without affixes—except for both SC performance measures, in which **morph.cx** performed slightly worse than **morph.cx.ax**. Thus, when excluding closed-class words in morphologically decomposed corpora, retaining affixes seems to improve performance. The same result emerged, however, when comparing decomposed factor combinations without closed-class exclusions—factor combinations with affixes outperformed those without (i.e., **morph.ax** vs. **morph**; **pos.morph.ax** vs. **pos.morph**; see Table 2)—except for SC percentage performance measures, on which **morph** performed worse than **morph.ax**. Overall, this suggests that a performance benefit may be associated with retaining affixes—and subsequently using them as context—in morphologically decomposed word–word co-occurrence semantic space models.

From these results, little can be said about whether affixes can provide some of the same semantic (via syntactical information) information provided by closed-class words, unless one assumes closed-class words contain *primarily* syntactic information. In that case, it could be claimed the results reported here provide support for the claim that affixes retain syntactic information in morphologically decomposed word–word co-occurrence semantic space models, as was demonstrated by their inclusion resulting in consistently recapturing some of the performance lost by excluding the syntactic information contained in closed-class words.

Furthermore, in all tasks and measures contrasting POS tagging and closed-class word exclusions (i.e., **pos.cx** vs. **pos**; **pos.morph.cx** vs. **pos.morph**; **pos.morph.cx.ax** vs. **pos.morph.ax**; see Table 2), factor combinations with POS tagging and *without* closed-class exclusions outperformed the equivalent factor combination *with* closed-class exclusions—except in the LD task, in which **pos.morph** performed slightly worse than **pos.morph.cx**. Indeed, POS-tagged factor combinations without closed-class exclusions performed much

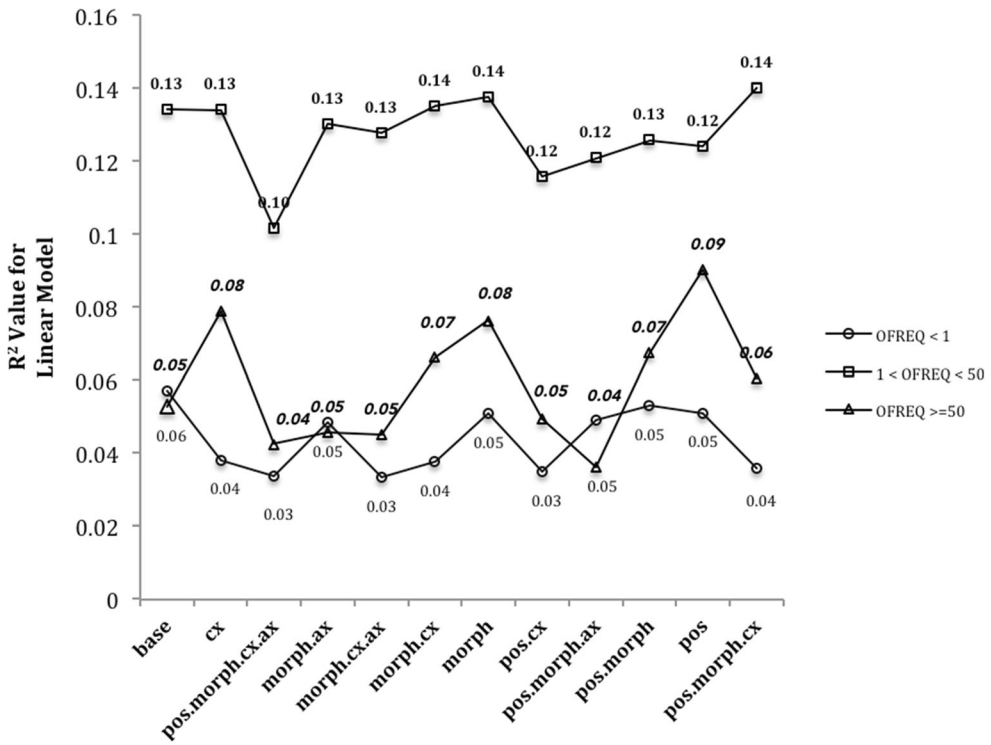

Fig. 7 Comparisons of variances accounted for (R^2) between lexical decision task regression models and the baseline, by orthographic frequency (OFREQ) bins (i.e., individual regression models were built using only those data with the specified OFREQ values). OFREQ values shown represent the numbers of target word occurrences per million words of text

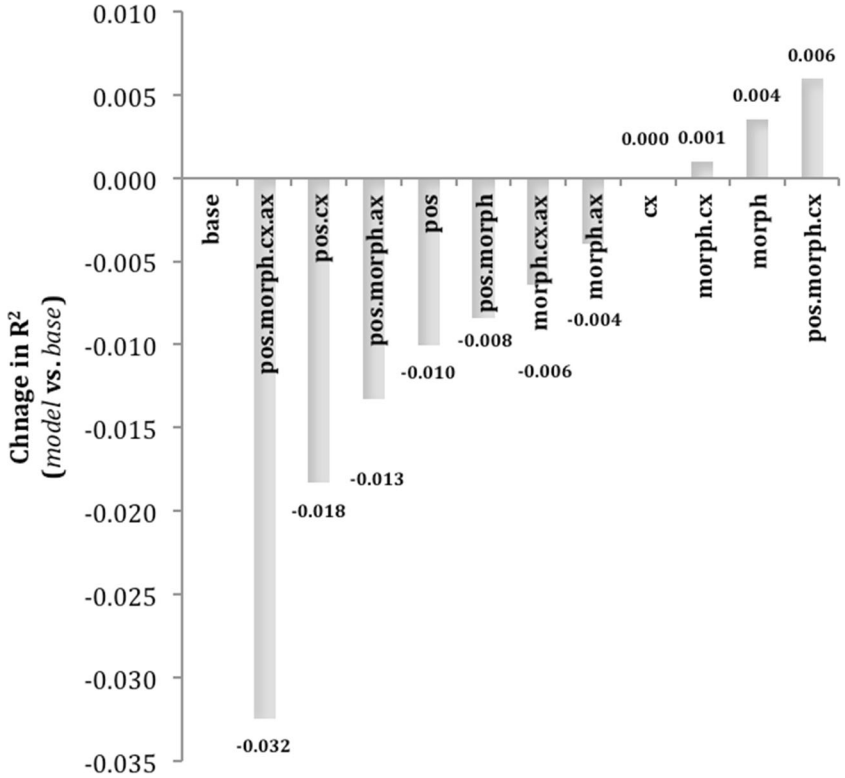

Fig. 8 Change in variances accounted for (ΔR^2) between lexical decision task regression models and the base model, where all models were built using only those data for which the target word's orthographic frequency was moderate (i.e., between 1 and 50 occurrences per million, representing the majority of the lexical decision target words)

better than those with exclusions; the latter were among the worst performers in all tasks and measures (note that in three of the six tasks and measures, **pos.morph.cx.ax** is the worst performer). This implies that POS tagging and closed-class words do not provide the same syntactic information in these factor combinations, and closed-class exclusions appear to worsen performance in POS-tagged factor combinations to the same (relative) extent as they do in non-POS-tagged factor combinations. This seems to weaken the claim that closed-class words supply *primarily* syntactic information in word–word co-occurrence semantic space models (see, e.g., Rohde et al., 2006).

This effect could instead be the result of the high frequency of closed-class words. HiDEx selects its context dimensions from the highest-frequency words/tokens, and there is a putative performance advantage of frequency-sorted contexts in some, but not all, co-occurrence models (for positive evidence regarding HAL-based models—like HiDEx—see Bullinaria & Levy, 2007, 2012). In other words, it seems that semantic information content can be richer in those context dimensions derived from higher-frequency words/tokens, and because closed-class words (and all other non-closed-class words on our stop lists) comprise the highest-frequency tokens in the corpus, removing closed-class words would then remove some of the richest information available in the co-occurrence statistics. Therefore, the assertions made regarding closed-class words may be unwarranted, and the observed effects may simply be artifacts of the co-occurrence implementation selected.

Performance impact of morphological decomposition

We also considered it likely that, by using a morphologically decomposed corpus, in which all contextual information for each inflected form of a word is combined into a single, monomorphemic lexical stem vector, a richer semantic representation of that word could be produced; or, alternatively, this could introduce excessive noise into the context and reduce semantic information content. Either way, morphological decomposition was expected to have a significant impact on performance.

Baayen and colleagues (2011) provide an interesting counterview to this expectation, having developed a highly effective computational language comprehension model that completely ignores morphology, as traditionally conceived. Their naive discriminant learning model (NDL) instead relies exclusively on orthographic *n*-grams (i.e., both subword and submorpheme) as the basic elements onto which meaning is mapped. In this model, traditional morphemes (and, indeed, words themselves) are simply probable sequences of letters. Taken as a model of human language processing, NDL would be entirely unaffected by morphological decomposition (or, indeed, any kind of systematic decomposition), so long as the morphemes were retained, as they play an important role

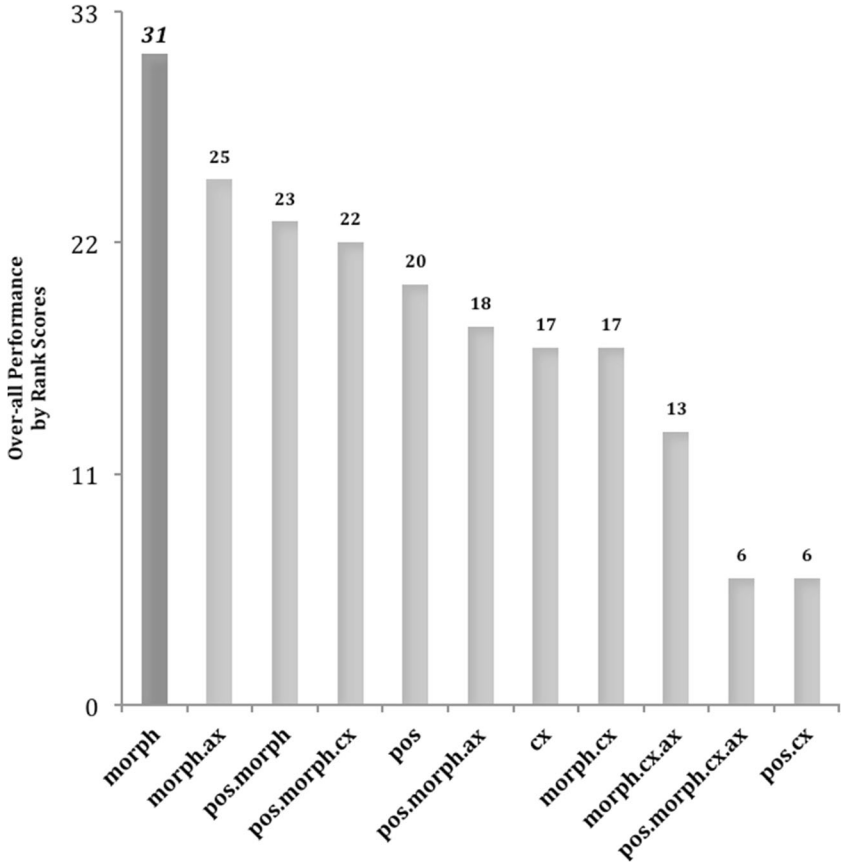

Fig. 9 Total performance scores, using only those tasks with the most reliable performance measures: Distance comparison and semantic categorization beta regressions, and lexical decision regressions only for words with moderate orthographic frequency (out of 33 max)

in delimiting the probabilities of subsequent letters that NDL is computing. Since it assumes that the mapping from letters to semantics is entirely a function of the probabilities of sequences of letters, NDL would predict that there should be no advantage for semantic access as a function of morphological decomposition.

Other studies have shown positive performance results for morphological decomposition. Krovetz (1993), for example, showed that decomposition via word stemming improved performance in various tasks by up to 45.3%. Hull (1996) similarly concluded that stemming is almost always beneficial, but, disagreeing with Krovetz, claimed the average improvement due to stemming is only 1% to 3%. Kiela and Clark (2014) found stemming improved performance in their models, however, they excluded closed-class words from all models, which potentially confounds their baseline performance making it difficult to compare their findings and ours.

It has also been claimed that morphological decomposition has, at best, no impact on performance, and often worsens it. Harman (1991) found that decomposition via word stemming provided no performance improvement, regardless of the stemming algorithm used. Bullinaria and Levy (2012) reported that morphological decomposition afforded no significant improvement to performance of word–word co-occurrence models. Using the same simple word-stemming approach taken by Bullinaria and Levy (2012), Landauer et al. (2007)—using a standard LSA implementation—likewise found that stemming offered no improvement to performance.

However, in all studies showing evidence against morphological decomposition's purported performance benefit, full morphological decomposition was not achieved. Decomposition in these studies was carried out by either simple word-stemming algorithms, or by using a standard lemmatized corpus. In contrast, the approach employed in this study—using PC-KIMMO and PrepCorpus—achieved approximately 99% accuracy in morphological decomposition and retained affixes for use as context in the models.

Both stemming and lemmatization reduce inflected word forms to a common base form—not necessarily to the word's morphological stem. The preponderance of stemming was carried out using Porter's (1980) algorithm, which employs a heuristic approach that only removes word suffixes and some derivational affixes. Lemmatization typically uses a predefined morphological analysis of words, but it is difficult to perform accurately for very large corpora because of its dependency on the word context. Neither of these approaches performs a full morphological decomposition, nor do they retain stripped affixes as context.

It might seem that stemming must necessarily fail in a co-occurrence model, since stemming removes semantic information. The word *runner* does not mean the same thing as the word *run*, but discarding the second morpheme in stemming *runner* would make the two words indistinguishable. However, the co-occurrence neighborhoods of words in nondecomposed models often include morphological variants of the same word (e.g., the first two neighbors of the word *work* are *working* and *works*), reflecting the fact that many morphological variants of the same root word share similar contexts. However, this is of course not always true; recall that (as cited above) Landauer and Dumais (2008) noted that "stemming often confabulates meanings" (p. 4356) in word vectors.

This study showed that decomposition while retaining the affixes appears to significantly improve performance in word–word co-occurrence semantic space models. Morphologically decomposed factor combinations (with no other manipulations) achieved the best overall performance, and outperformed baseline in all of tasks and measures (except in the—least reliable—SC percentage measure).

One possible explanation for this finding is that having word vectors with more context information (i.e., information from all inflected forms combine into one vector for the monomorphemic word stem) increases the semantic information contained therein. In support of this possibility, Rapp (2003) developed a relatively simple algorithmic machine translation approach to inducing context-dependent word sense (e.g., disambiguating homographs) from co-occurrence statistics. In doing so, Rapp demonstrated that word vectors in such models contain an aggregate representation of the underlying semantics, which implies content rich vectors (i.e., those collecting co-occurrence statistics for a word used in multiple contexts and senses, as affixes are) can provide better representations of a given word's semantic content. Similarly, Jing and Tzoukermann (1999) demonstrated that morphological information improved calculations of semantic relatedness between two words (i.e., by computing the distance between their vectors using their own model based on second-order, word–word co-occurrence statistics).

As well as aggregating information, and as we have noted briefly above, morphological decomposition provides more contexts for the root word. This potentially allows for fine tuning of the associate/semantic information its neighborhood contains. Consider the word discussed earlier, *computerization*, consisting of four morphemes *compute* + *er* + *ize* + *ation*. When it is entered as a single word, the root word *compute* does not benefit from exposure the context of computerization, though undoubtedly information about what compute means enters into contexts discussing computerization. When the word is decomposed, the word compute does benefit from the context in which computerization was used. In cases in which the affixed and unaffixed forms are both commonly used—such as many singular and pluralized nouns—the roots are likely to gain a very large increase in exposure to contexts.

We also note that, independently of whether words are decomposed or not when they are accessed, the semantic aggregation of related words happens naturally, easily, and early in real life. We know that running is the same thing whether *we ran*, *are going to run*, or *have been running*; we know that dogs do not change kind when they aggregate in numbers greater than the singular; and we know that computerization is about computers. We aggregate our experiences of the same thing, and humans can benefit from that aggregation linguistically, independently of whether or not it occurs as a result of morphological decomposition.

Context selection provides a third possible explanation for the observed improvement in performance of morphologically decomposed factor combinations. HiDEx selects its context dimensions by taking the N highest frequency words/tokens (where N is a user-defined parameter; N = 15,000 in this study). There is a putative performance advantage of frequency-sorted contexts in some, but not all, co-occurrence models. Bullinaria and Levy (2007, 2012) provide positive evidence for this being the case in HAL-based models, like HiDEx. In other words, it is claimed that semantic information content is richer in context dimensions derived from higher frequency words/tokens. A morphologically decomposed corpus will necessarily have more higher frequency words/tokens than a nondecomposed version of the same corpus. Multiple inflected word forms will each be decomposed to—and have their occurrences contribute to the frequency of—a single, common stem. Therefore, the positive results for morphological decomposition reported here might be the result of a systematic increase in the frequency of context dimensions. In other words, the observed effects may simply be artifacts of the co-occurrence implementation used (i.e., HiDEx). This suggests a direction for future study, specifically, investigating whether the same increase in performance resulting from full morphological decomposition is observed in word–word co-occurrence semantic space models that use variance, rather than frequency, selected context dimensions.[12]

These three reasons provide explanations for the decomposition advantage for root words without making reference to the information that is contained in the context dimensions for the affixes. Since affixation is often widely applicable to many radically different words (e.g., we can *computerize*, *demonize*, *vaporize*, etc.) affix vectors (like, e.g., +*ize*) must be

[12] Though frequency and variance in word-based context dimensions are highly correlated, there is evidence for differential performance between these two types of context selection processes (see, e.g., Levy & Bullinaria, 2001). For a more elaborate discussion of using and manipulating variance-selected context dimensions, please refer to Bullinaria and Levy (2012).

uninformative about their context, which is a melange of otherwise-unrelated contexts held together by a common affix. However, as we noted above, despite providing little useful information about their own contexts, affix context dimensions do provide useful context for their root word, since—for example, they do a little to group together all things that can be *+ized*, thus pushing nouns (things that be *+ized*) a little closer together and a little further from other classes of things (i.e., we cannot verb-ize or pronoun-ize). When aggregated together, the very common co-occurrence information from all morphemes may push words around a little in co-occurrence space, and that movement may—as our results suggest—help increase semantic differentiation and accuracy.[13]

It has been found that stemming greatly improves some individual queries and severely degrades others (Krovetz, 1993), and it has been further suggested that this tends to conceal any improvement in overall performance results (Hull, 1996). More accurate queries result from morphological variants being semantically related; in other cases, in which variants are not semantically related, word stemming introduces noise (Church, 1995; Krovetz, 1993). As another direction for future research, this problem could possibly be addressed—and stronger, more consistent performance impacts observed—through correlating morphologically related word vectors, using the results to distinguish the cases in which decomposition helps (i.e., high correlations) from those in which it does not, and morphologically decomposing/stemming only positive cases (e.g., Xu & Croft, 1996).

A final word on methodological limitations

Our findings provide support for the claim that sublexical information—specifically word morphology—plays a role in lexical semantic processing. Interpretation of these findings, however, ought to be tempered by recognizing some potential limitations imposed by our methodology.

The present study focused on word–word corpus-based semantic space models. Employing full morphological decomposition in other distributional models of semantics—for example, the word–document paradigm employed by LSA—would lend insight into the generalizability of our findings. More, we made use of a single corpus of text for each of the 12 factor combinations considered. Extending this work to explore the performance impact of full morphological decomposition across different corpora of text would also better inform claims of generalizability.

As they were implemented in this study, our morphologically decomposed factor combinations—with or without affixes included in the co-occurrence statistics—could not deal with distances between multimorphemic words in a practical manner. If, for example, we wanted to compare distances in the semantic space between *unhappiness* and *happiness*, in decomposed models we end up comparing *happy* to *happy*. In a way, however, that is exactly the point of these models; they simply ignore those differences that are "just" due to affixation. In morphologically decomposed models in which affixes are retained, it may be feasible to construct representations for multimorphemic words such as (e.g.) *unhappiness* and *happiness*, but in such a case these words would both be represented by multiple vectors rather than a single vector. The approach for doing so, however, was considered beyond the scope of this study, and note that we disregarded inflected forms of targets word in comparison tasks (see the Methods section). Readers interested in constructing representations for multimorphemic words are encouraged to consider some more recent work being done with vector composition models, operating at the level of morphemes (e.g., Lazaridou, Marelli, Zamparelli, & Baroni, 2013; Luong, Socher, & Manning, 2013) and phrases (e.g., Mitchell & Lapata, 2010).

Also, as was mentioned in the Method section, compound words (e.g., *sunshine*, *grandmother*, *scarecrow*, etc.) were not decomposed in this study. This decision was made on the basis of previous work supporting dual-processing models of multimorphemic words, wherein multimorphemic words—and particularly compound words—are purported to afford lexical access, and maintain semantic representation in memory, without requiring obligatory morphological decomposition (e.g., Bybee, 1995; Elman, 2004; Kuperman, Schreuder, Bertram, & Baayen, 2009; Rubin, Becker, & Freeman, 1979; Seidenberg & Gonnerman, 2000). We acknowledge, however, that this is a contentious claim, and that evidence has also been provided against these dual-processing models, asserting the obligatory decomposition of multimorphemic words (e.g., Butterworth, 1983; Caramazza, 1997; Fiorentino & Poeppel, 2007; Marslen-Wilson & Zhou, 1999; Stockhall & Marantz, 2006; Taft, 2004).

Author note J.K. completed this article as his undergraduate honors thesis at the Department of Psychology, University of Alberta. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to the second author.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 234–254. doi:10.1037/0278-7393.18.2.234

[13] A similar argument could be made for certain closed-class words, which may help account for the performance advantaged noted in this study for retaining closed-class words as context dimensions.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review, 118*(3), 438.

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

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation, 43,* 209–226.

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology, 80*(3, Pt. 2), 1–46. doi:10.1037/h0027577

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review, 8,* 531–544. doi:10.3758/BF03196189

Bullinaria, J. A. (2008). Semantic categorization using simple word co-occurrence statistics. In M. Baroni, S. Evert, & A. Lenci (Eds.), *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics* (pp. 1–8). Hamburg, Germany: ESSLLI.

Bullinaria, J. A. (2013). *Corpus derived semantic representations*. Retrieved from www.cs.bham.ac.uk/~jxb/corpus.html

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39,* 510–526. doi:10.3758/BF03193020

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods, 44,* 890–907. doi:10.3758/s13428-011-0183-8

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers, 30,* 188–198. doi:10.3758/BF03200643

Butterworth, B. (1983). Lexical representation. In B. Butterworth (Ed.), *Language production* (pp. 257–294). New York, NY: Academic Press.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10,* 425–455.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology, 14,* 177–208.

Church, K. W. (1995). One term or two?. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 310–318). ACM.

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, *34*, 1–24. Retrieved from www.jstatsoft.org/v34/i02/

Damerau, F. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management, 29,* 433–447.

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

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences, 8,* 301–306. doi:10.1016/j.tics.2004.05.003

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31,* 799–815.

Finch, S. P., & Chater, N. (1992). Bootstrapping syntactic categories. In J. K. Kruschke (Ed.), *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 820–825). Hillsdale, NJ: Erlbaum.

Fiorentino, R., & Poeppel, D. (2007). Compound words and structure in the lexicon. *Language and Cognitive Processes, 22,* 953–1000.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. In J. R. Firth (Ed.), *Studies in linguistic analysis* (pp. 1–32). Oxford, UK: Blackwell.

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika, 40,* 237–264.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language, 29,* 228–244. doi:10.1016/0749-596X(90)90074-A

Harman, D. (1991). How effective is suffixing? *JASIS, 42,* 7–15.

Hull, D. A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science, 47,* 70–84.

Jing, H., & Tzoukermann, E. (1999). Information retrieval based on context distance and morphology. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 90–96). ACM.

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

Kiela, D., & Clark, S. (2014, April). A systematic study of semantic vector space model parameters. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) at EACL* (pp. 21–30).

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York, NY: Cambridge University Press.

Kintsch, W. (2001). Predication. *Cognitive Science, 25,* 173–202. doi:10.1207/s15516709cog2502_1

Koskenniemi, K. (1984). A general computational model for word-form recognition and production. In *Proceedings of the 10th international conference on Computational Linguistics* (pp. 178–181). Association for Computational Linguistics.

Krovetz, R. (1993). Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 191–202). ACM.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. H. (2009). Reading of multimorphemic Dutch compounds: Towards a multiple route model of lexical processing. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 876–895.

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,* 211–240. doi:10.1037/0033-295X.104.2.211

Landauer, T. K., & Dumais, S. T. (2008). Latent semantic analysis. *Scholarpedia, 3,* 4356.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284. doi:10.1080/01638539809545028

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Lapesa, G., & Evert, S. (2013). Evaluating neighbor rank and distance measures as predictors of semantic priming. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics* (CMCL 2013) (pp. 66–74).

Lazaridou, A., Marelli, M., Zamparelli, R., & Baroni, M. (2013). Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics. In *ACL, 1,* 1517–1526.

Levy, J. P., & Bullinaria, J. A. (2001). Learning lexical properties from word usage patterns: Which context words should be used? In R. F. French & J. P. Sougne (Eds.), *Connectionist models of learning, development and evolution* (pp. 273–282). Heidelberg, Germany: Springer.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28,* 203–208. doi:10.3758/BF03204766

Luong, M. T., Socher, R., & Manning, C. D. (2013). Better word representations with recursive neural networks for morphology. *CoNLL-2013, 104*.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: Freeman.

Marslen-Wilson, W., & Zhou, X. (1999). Abstractness, allomorphy, and lexical architecture. *Language and Cognitive Processes, 14,* 321–352. doi:10.1080/016909699386257

McCann, R. S., Besner, D., & Davelaar, E. (1988). Word recognition and identification: Do word-frequency effects reflect lexical access? *Journal of Experimental Psychology: Human Perception and Performance, 14,* 693–706. doi:10.1037/0096-1523.14.4.693

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science, 34,* 1388–1429. doi:10.1111/j.1551-6709.2010.01106.x

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin, 49,* 197–237. doi:10.1037/h0055737

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.

Patel, M., Bullinaria, J. A., & Levy, J. P. (1997). Extracting semantic representations from large text corpora. In J. A. Bullinaria, D. W. Glasspool, & G. Houghton (Eds.), *Fourth Neural Computation and Psychology Workshop: Connectionist representations* (pp. 199–212). London, UK: Springer.

Perfetti, C. A. (1998). The limits of co-occurrence: Tools and theories in language research. *Discourse Processes, 25,* 363–377.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14,* 130–137.

R Development Core Team (2013). *R: A language and environment for statistical computing* [Software environment]. Vienna, Austria: R Foundation for Statistical Computing. URL www.R-project.org/

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Ninth Machine Translation Summit* (pp. 315–322).

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

Rubin, G. S., Becker, C. A., & Freeman, R. H. (1979). Morphological structure and its effect on visual word recognition. *Journal of Verbal Learning and Verbal Behavior, 18,* 757–767.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 1–17. doi:10.1037/0096-1523.3.1.1

Schütze, H. (1993). Word space. In S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.), *Advances in neural information processing systems* (pp. 895–902). San Mateo, CA: Morgan Kaufmann.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization*. Norwood, NJ: Ablex.

Seidenberg, M. S., & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences, 4,* 353–361.

Sequence Publishing (2014). *Function words* [Data file]. Retrieved from www.sequencepublishing.com/academic.html

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods, 38,* 190–195. doi:10.3758/BF03192768

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods, 42,* 393–413. doi:10.3758/BRM.42.2.393

Shaoul, C., & Westbury, C. (2012). HiDEx: The high dimensional explorer. In P. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation and resolution* (pp. 230–246). Hershey, PA: Information Science Reference. doi:10.4018/978-1-60960-741-8.ch013

SIL International. (1997). *PC-KIMMO* (Version 2.0) [Software]. Available from http://www.sil.org/computing/catalog/pc-kimmo.html

Smith, A. E., & Humphreys, M. S. (2006). Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping. *Behavior Research Methods, 38,* 262–279. doi:10.3758/BF03192778

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods, 11,* 54–71. doi:10.1037/1082-989X.11.1.54

Solomyak, O., & Marantz, A. (2010). Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience, 22,* 2042–2057.

Sproat, R. (1991). Review of "PC-KIMMO: A two-level processor for morphological analysis" by Evan L. Antworth. *Computational Linguistics, 17,* 229–231.

Stockhall, L., & Marantz, A. (2006). A single route, full decomposition model of morphological complexity: MEG evidence. *Mental Lexicon, 1,* 83–123.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology, 57A,* 745–765. doi:10.1080/02724980343000477

Turney, P. (2006). Similarity of semantic relations. *Computational Linguistics, 32,* 379–416.

Tysto. (2012). *Comprehensive list of American and British spelling differences* [Data file]. Retrieved from www.tysto.com/uk-us-spelling-list.html

Wikipedia. (2014). *List of spelling variants* [Data file]. Retrieved from http://en.wikipedia.org/wiki/Wikipedia:List_of_spelling_variants

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. In *SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 4–11). ACM.