



# Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging

Chris Westbury<sup>1</sup> · Geoff Hollis<sup>1</sup>

Published online: 13 September 2018  
© Psychonomic Society, Inc. 2018

## Abstract

Co-occurrence models have been of considerable interest to psychologists because they are built on very simple functionality. This is particularly clear in the case of *prediction models*, such as the *continuous skip-gram model* introduced in Mikolov, Chen, Corrado, and Dean (2013), because these models depend on functionality closely related to the simple Rescorla–Wagner model of *discriminant learning* in nonhuman animals (Rescorla & Wagner, 1972), which has a rich history within psychology as a model of many animal learning processes. We replicate and extend earlier work showing that it is possible to extract accurate information about syntactic category and morphological family membership directly from patterns of word co-occurrence, and provide evidence from four experiments showing that this information predicts human reaction times and accuracy for class membership decisions.

**Keywords** Semantics · Morphology · Co-occurrence models · Word2vec · Part-of-speech tagging

Traditionally, language has been conceived as having several quasi-independent components: phonology/orthography, morphology, semantics, and syntax. Statistical processing models of language can blur the apparently clear distinctions between these components, by, for example, taking a purely statistical approach to explaining variations in human behavior related to morphology (e.g., Baayen, Milin, Đurđević, Hendrix, & Marelli, 2011) or variation in syntactic judgments (e.g., Roberts & Chater, 2008). The second-order co-occurrence statistics used in distributional models have previously been shown to capture a great deal of semantic information (e.g., Landauer & Dumais, 1997; Mandera, Keuleers, & Brysbaert, 2017; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Shaoul & Westbury, 2008). It has also been demonstrated that these distributional models encode syntactic information (e.g., Abka, 2016; Baroni & Zamparelli, 2010; Burgess & Lund, 2000; Cotterell & Schütze, 2015; Drozd, Gladkova, & Matsuoka, 2016; Gladkova, Drozd, &

Matsuoka, 2016; Lazaridou, Marelli, Zamparelli, & Baroni, 2013; Lin, Ammar, Dyer, & Levin, 2015; Ling, Dyer, Black, & Trancoso, 2015; Marelli & Baroni, 2015; Mikolov, Yih, & Zweig, 2013). In this article, we compile further evidence showing that second-order co-occurrence statistics can be used to classify words in a large number of different word categories using simple methods closely drawn from semantic analyses; demonstrate that it is possible to conceptualize part-of-speech category membership as a continuous measure (a weighted sum of the principal components of co-occurrence variance); and provide strong evidence of the behavioral relevance of this continuous measure of word class membership.

## Background

There is a large literature on computational part-of-speech tagging, which has long ago reached high levels of accuracy (i.e., 96.6% in Ratnaparkhi, 1996). As psychologists, our interest is not in computational approaches to part-of-speech tagging per se, but rather in computational approaches to part-of-speech tagging whose psychological plausibility and quantified outcome measures have the potential to shed light on linguistic processing. In this context, distributional models have been of considerable interest to psychologists by virtue

✉ Chris Westbury  
chrisw@ualberta.ca

<sup>1</sup> Department of Psychology, University of Alberta, P220 Biological Sciences Building, Edmonton, Alberta T6G 2E9, Canada

of the fact that they are built on very simple functionality. This is particularly clear in the case of *prediction models*, such as the *continuous skip-gram model* introduced in Mikolov, Chen, Corrado, and Dean (2013; see also Baroni, Dinu, & Kruszewski, 2014; Mikolov, Sutskever, et al., 2013). Instead of counting how often words occurred in proximity to each other, as previous co-occurrence models of semantics had done (e.g., Durda & Buchanan, 2008; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996; Shaoul & Westbury, 2008, 2010, 2011), Mikolov, Chen, et al.'s model used a simple three-layer neural network with back-propagation to try to predict the surrounding context of each target word (e.g., the two words before and after), adjusting the weights on that word's vector in order to optimize the success of the network's prediction.

Skip-gram models were designed to model semantics, by making it possible to quantify the relationship between two word meanings as the cosine similarity of the vectors representing those two words. They are very good at capturing word-to-word semantic relationships, and multi-word to word semantic relationships. For example, in our matrix of 78,278 words, the ten words with the most similar vectors to the word *dog* are *dogs*, *puppy*, *pooch*, *cat*, *golden\_retriever*, *beagle*, *pup*, *canines*, *pet*, and *schnauzer*. If instead we average together the vector for a set of ten random mammals (*dog*, *aardvark*, *antelope*, *pig*, *giraffe*, *mouse*, *moose*, *jaguar*, *panda*, and *bettong* [the little-known Tasmanian rat-kangaroo]), then (ignoring the ten names in that category-defining list, some of which are echoed back) the closest neighbors are *rabbit*, *rhino*, *cat*, *critter*, and *bobcat*. The sixteenth closest neighbor is the word *mammal*. The first 100 neighbors include 82 mammal names. This ability of averaged vectors to define semantic categories is very flexible and not limited to natural categories such as *mammal*. If we average together the vectors for the 17 country names that appear in our dictionary from a randomly selected 25 country names, the first 100 neighbors of the averaged vector include 94 country names, plus the words *Africa*, *Caribbean*, *Colombian*, *countries*, *Peruvian*, and *Venezuelan*.

As a model of word meaning, the skip-gram model is conceptually more plausible than (although functionally very similar to) earlier count models, since skip-gram builds directly on the Rescorla–Wagner model of discriminant learning in nonhuman animals (Rescorla & Wagner, 1972). Although the Rescorla–Wagner model has some limitations, it does capture a wide range of key findings in the classical conditioning literature (see Miller, Barnet, & Grahame, 1995, who also point out some of the limitations of the model) and has, as a result, been one of the most influential models of animal learning (Siegel & Allan, 1996). It also has the virtue of being “the simplest possible learning rule in which prediction error plays a central role” (Milin, Feldman, Ramscar, Hendrix, & Baayen, 2017, p. 8). The Rescorla–Wagner model conceptualizes learning as a process of reducing uncertainty or (the same thing said another

way) maximizing predictability from an animal's point of view (Rescorla, 1988). Under the model, discrete cues that do not allow predictions relevant to the learning environment (i.e., *uninformative cues*) are given lower-magnitude weights, whereas cues that do allow relevant predictions (*informative cues*) are given higher-magnitude weights, thereby decreasing the discrepancy between what is predicted and what is experienced. The skip-gram model was explicitly designed to minimize the prediction error between a cue (the target word) and a linguistic environment (the words surrounding the target word). Its close relationship to the Rescorla–Wagner model is explicit, since the delta rule (Rosenblatt, 1957) it uses to update its vector values is mathematically equivalent to the update rule in the Rescorla–Wagner model (Sutton and Barto, 1998). Rescorla (2008) himself noted this, writing that the Rescorla–Wagner rule “is essentially identical to the learning algorithm of Widrow and [sic] Hoff (1960) which closely corresponds to the delta rule implemented in many connectionist networks” (paragraph 7). Skip-gram is not a perfect instantiation of the Rescorla–Wagner model, since it uses a nonlinear activation function and includes a hidden layer (see the discussions in Mandera et al., 2017), but it is conceptually very similar.

Crucially, both positive evidence (i.e., support for a hypothesis based on the presence of a cue) and negative evidence (i.e., support for a hypothesis based on the absence of a cue) play vital roles in the Rescorla–Wagner model. The model is deliberately structured in such a way that negative evidence reduces the association between a cue and a *nonoccurring* expected outcome. This makes the model of particular interest to psychologists due to its relevance to the *poverty-of-the-stimulus* argument (Chomsky, 1980), sometimes called the *logical problem of language acquisition* [LPLA] (see Baker, 1979; Pinker, 1984, 1989). The LPLA is that the linguistic environment to which children are exposed is too impoverished to support language learning, since children are rarely provided with explicit negative feedback about language errors. Prediction models of learning like the Rescorla–Wagner model address this problem, because they build in negative evidence in the form of an organism making, and correcting for, erroneous predictions about its own environment.

We note that the use of the term *negative evidence* when talking about the LPLA typically refers to feedback about the incorrectness of an utterance—for example, through explicit correction from parents (though as we note above, this correction need not come from an external source—it is available when a model's predictions do not match the observations). This use of the term is in contrast with how it is used within the animal learning literature, where it means support for a hypothesis (e.g., a specific grammar; a specific cue–outcome relationship) based on the nonoccurrence of a stimulus or outcome. This difference in terminology may reflect a

misinterpretation of learning theory by researchers working on the LPLA. As Ramscar, Dye, and McCauley (2013) point out, the logic of the LPLA falls apart when reframed as an animal learning problem:

The discovery that animals are perfectly capable of learning about predictive relationships even when they have no explicit access to the locus of their predictions contrasts with a critical assumption in the LPLA—and much of the language learning literature—that learned inferences can only be unlearned when explicit correction is provided (Baker 1979, Brown & Hanlon 1970, Marcus 1993, Marcus et al. 1992, Pinker 1984, 1989, 2004) [sic]. If the logic of the LPLA were applied to rat learning, it would predict that rats could only learn about the relationship between a tone and an absent shock if they were provided with additional, explicit information about this relationship. Rescorla's—and countless other—experiments make clear that, for many species of animals, at least, this prediction is simply false. (p. 766)

As they go on to note, a great deal of evidence shows that even very young children are sensitive to the statistical structure of language (Ramscar, Dye, & Klein, 2013; Ramscar, Dye, Popick, & O'Donnell-McCarthy, 2011; Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010; Saffran, 2001; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999). Given this evidence, and the reasonable assumption that human children are at least as capable of learning about the statistics of discriminant cues as rats are, there has been growing interest recently in conceptualizing language learning as a problem of discriminant learning (Baayen et al., 2011; Mandera et al., 2017; Milin et al., 2017; Ramscar, Dye, & McCauley, 2013). Under this conception, words (or possibly some other linguistic unit, as in Baayen et al., 2011) are cues used for predicting context. Language learners are able to use both positive evidence (correct prediction of a word's context) and negative evidence (nonoccurrence of an expected context) to build a statistical model of language.

Ramscar and his colleagues have conducted several experiments on children that were explicitly designed to test (and that did find evidence to confirm) the claim that discriminant learning is relevant to language learning (i.e., Arnon & Ramscar, 2012; Ramscar, Dye, Gustafson, & Klein, 2013; Ramscar, Dye, & Klein, 2013; Ramscar, Dye, & McCauley, 2013; Ramscar et al., 2011; Ramscar et al., 2010). For example, discriminant learning predicts that the order in which (the same) relationships are learned will impact on how well those relationships are learned, since prior learning will block later learning if the prior learning suggests that that later-learned relations are redundant. Arnon and Ramscar tested this

prediction in a study of the acquisition of gendered articles in an artificial language. They noted that when a noun occurs with its referent object (but without an article), the association between the noun and object will strengthen at the expense of the article. By the same token, when an article occurs without a particular noun, the relationship between that article and that noun will weaken, because the occurrence is negative evidence of an association between the two. They tested the prediction that learning smaller semantic units (nouns) in isolation should therefore make it more difficult to learn article–noun associations, since the association will be blocked by prior learning of the nonassociation. They had two experimental groups that were exposed to exactly the same information, but in a different order. The noun-first group was first exposed to nouns (smaller units first) and then to full sentences containing article–noun pairings. The sequence-first group (larger units first) was exposed first to the sentences and then to the nouns. In testing, as predicted by discriminative learning, the participants in the sequence-first group learned the artificial gendered articles better than the noun-first group, as tested both by forced-choice recall and by production. A computer simulation using the Rescorla–Wagner model showed the same effects.

The close similarity between the skip-gram model and discriminant learning makes it possible to conceptualize lexical semantics as an (admittedly rather spectacular) extension of a well-recognized simpler psychological function that is ubiquitous across the animal kingdom. This conceptualization directly addresses the LPLA, which was founded on the notion that language learners do not get negative feedback. When we think of language as a prediction problem, every word provides negative or positive feedback, since a language user either correctly predicts the context in which a word will appear or she does not.

In one of the first articles on co-occurrence models of language, Burgess and Lund (2000) included a proof of concept showing that multidimensional scaling was able to use co-occurrence information from their HAL model to correctly separate just 11 words into three part-of-speech categories. More recent work (e.g., Drozd et al., 2016; Gladkova et al., 2016; Mikolov, Yih, & Zweig, 2013) has shown that word vectors in co-occurrence models captured syntactic regularities, by showing that the model could produce inflected versions of some base words. Many of these articles used the *vector offset method*, which depends on drawing an analogy (by comparing vector differences) to a known pair. A frequently cited semantic example (from Mikolov, Chen, et al., 2013) is that it is possible to solve the analogy *king:man :: X:queen* using simple vector operations, because the vector derived by subtracting *king* – *man* + *woman* is closely similar to the vector for *queen*. With a very large test set covering 40 semantic and morphological categories (and after eliminating homonymous words, whose disambiguation we consider

below in Exps. 3 and 4), Gladkova, Drozd, and Matsuoka (2016) were able to find the correct analogies 28.5% of the time, on average. The best performance was seen in predicting the plural form of a singular noun, at which the model succeeded about 80% of the time (exact values were not provided in the article and are here estimated from a graph). Note that these test sets scored results across many individual analogies. The method we discuss in this article can be conceptualized as a way of estimating the global offset for an entire category of interest directly, instead of computing offsets individually for individual analogies within that category of interest.

Building on somewhat related work on adjective–noun constructions in Baroni and Zamparelli (2010), Marelli and Baroni (2015) represented affixed words in their 20,000 word matrix as vectors of length 350 for the stem multiplied by a  $350 \times 350$  matrix (i.e., a weighted linear function) for the affix. The affix matrix was estimated by training it on at least 50 stem/derived words pairs, so that multiplication of the stem by the matrix would yield new vectors of length 350 that were as close as possible to the vectors for the derived words in the training set. Their method involves a great deal of computation. Marelli and Baroni noted that much simpler models are conceivable, writing that

It may be tempting to conceive a semantic system populated by full-form meanings (i.e., separate representations for *run*, *runner*, *homerun*) and explain alleged morphological effects as by-products of semantic and formal similarity, and/or lexical links between related whole-word representations. This solution permits dealing with the idiosyncratic semantics characterizing (to different degrees) nearly all complex words. It can also handle cases where a complex form contains a reasonably transparent affix meaning but the stem is not a word: *grocer* clearly displays the agentive sense of *-er*, but *to groce* is not a verb, so the noun cannot be derived compositionally. (pp. 9–10).

Ultimately, Marelli and Baroni (2015) rejected this semantic approach to morphology on the grounds that “holistic meanings by themselves fall short in explaining the surprising productivity of morphological systems” (p. 10). In this article, we challenge this conclusion by demonstrating how much relevant work can be done using precisely the method that Marelli and Baroni rejected. We show that it is possible to extract a great deal of accurate and highly specific morphological and part-of-speech information directly from patterns of whole-word co-occurrence, using vector addition, without paying any attention to the formal structure of the words. We also discuss the ability of such a model to address the issue of morphological productivity, using the same simple vector operations.

## Part 1: Proof of concept

Hollis and Westbury (2016) used principal component analysis (PCA) to analyze a  $12,344 \text{ Word} \times 300 \text{ Feature}$  skip-gram matrix, in an attempt to identify the principal components of lexical semantics (following the work of Osgood, Suci, & Tannenbaum, 1957/1978). Here we build on that work by using a skip-gram matrix of size  $78,278 \text{ Words} \times 300 \text{ Features}$ , on which we have also performed PCA. This is a subset of the pretrained skip-gram matrix released by Google (<https://github.com/mmihaltz/word2vec-GoogleNews-vectors>), built from a three-billion-word Google News Corpus. We do not claim that the PCA is a theoretically important transformation and have previously shown (in unpublished work) that the cosine distances between words are nearly identical in both the transformed and untransformed matrices.

## Method

Here we extend the idea of defining semantic categories using average vectors from a semantic category, as outlined above. We average vectors to identify part-of-speech and morphological families. We will use the same simple method for estimating category membership throughout this article. In every case, we start with 100 randomly selected, human-verified exemplars of the category of interest. We compute the cosine distance from the vector containing the 300 average principal component (PC) values of those 100 exemplars (which we refer to as the *category-defining vector*, or *CDV*) as the measure for defining and quantifying category membership.

We note in advance that our classifications will always include the 100 randomly selected words that made up the sample we used to define each class’s CDV. This apparent circularity is not in any way problematic. We do not believe there is any theoretical or methodological issue that depends on whether we generate our models using 100 words or every English word in the category of interest. Our primary goal here is to demonstrate that there is sufficient information in the co-occurrence matrix (as a proxy for linguistic experience) to define and quantify graded membership in those categories. We demonstrate unambiguously that morphological and syntactic categories can be derived and graded for centrality solely from the statistical information encoded from word co-occurrence. The interesting fact that this is true does not depend in any strong way on how many words we use to show it to be true. We chose to use a small number (i.e., 100) in order to be able to clearly validate our method by extrapolating it to the entire dictionary. However, we suspect a larger number might make for even more robust models than the (already) robust models we present here. We will compare different word sets in the experimental section below and show that they produce nearly identical CDVs.

We report two measures of the correct classification rate, one for the first 1,000 words in the dictionary, ordered by similarity to the CDV when there are at least 1,000 members in the class of interest, and one for the first decile of the dictionary (i.e., for the 7,828 words with vectors most similar to the CDV, when we are using the full dictionary, or for 4,743 words, when we are using the subset of the dictionary that is part-of-speech-tagged). If the classification were completely at chance, 1/10 of the classified words should be found among the first decile of the dictionary. The further the increase from this chance distribution, the better the classification. We can quantify this by computing a chi-square test over the observed versus predicted (chance) occurrence of the words from the category of interest among the first 1,000 words and the first decile.

For the 1,000 words closest to the CDV, we verified the classification by hand—that is, we checked every word classified as belonging to a family to confirm that it did belong to that family. These reported classification rates over the nearest 1,000 words to the CDV are thus a “gold standard.”

When counting relevant occurrences in the first decile (and graphing their distribution across the entire dictionary), for morphologically defined categories, the correct classification rate is scored by form-matching (i.e., by counting how many words begin or end with the affix of interest, independently of their morphological status), whereas the classification is carried out semantically. Concretely, a word like *mechanic* will be classified as a miss for the morphological family NOUN+*er*. Though this indeed is an error morphologically, *semantically* it is quite correct: a mechanic does belong to the same class of actors as a polisher, an installer, a crimper, a writer, a logger, and so on. Moreover, the classifications by affix also count words that happen to include the letters of an affix, whether or not they are multimorphemic (e.g., the word *hic* is classified with the words suffixed with *-ic*; the words *Italy* and *imply* are classified with words suffixed with *-ly*). This also means our decile estimates (though not the hand-checked values from the 1,000 words closest to the CDV) will be slightly inaccurate. The classification rate nevertheless allows us to set a lower bound on morphological family classification accuracy. As we will see, that lower bound is often very high.

We also show below that we can identify a part of speech without relying on string resemblance due to sharing an affix. When we consider part-of-speech-defined categories, the correct classification rate was scored using a part-of-speech dictionary, which we describe below.

There are of course a great many morphemes in English. Here we discuss a few in detail. We present data about many more in summary form in Table 1.

## Results

**Suffix –ed** We begin with the past-tense-forming morpheme *-ed*, as in *canned*, *placed*, and *created*. The dictionary contains

6,965 words that end with *ed*, of which the vast majority are past tense verbs. As is shown in Fig. 1, distance from the CDV derived by averaging 100 randomly selected words ending with *ed* is a very good predictor of ending with the morpheme *-ed*. In all, 72.9% of final-*ed* words are in the first decile of the dictionary (as contrasted to the 10% expected by chance) after sorting it by similarity to the class-defining vector [ $\chi^2(1) = 5,667.0, p \approx 0$ ].<sup>1</sup>

As outlined above, we examined the first 1,000 words manually. Of the 76 words (7.6%) that were not past tense verbs ending in *ed* [ $\chi^2(1) = 742.24, p = 9.69\text{E-}164$ ], 51 (5.1%) were irregular past tense verbs (e.g., *outshone*, *brought*, *ran*). All of the remaining 25 words were past participles (e.g., *gone*, *overtaken*, *withdrawn*).

We stress here the obvious but central point: The verbs being picked out by the category-defining vector are not *semantically* similar in the usual sense of the word. They do not have the same *meaning*. The similarity that is captured by the CDV is derived entirely from the fact that they are similar in one very abstract way: that they all refer to the past (or, more concretely, they share common neighbors, as we will demonstrate below).

**Morpheme –s** The second morpheme we consider is word-final *s*, which we also used in our experiments described below. This suffix is of interest since it is ambiguous, being used both to pluralize nouns (*cats*, *diamonds*, *pizzas*) and to mark verbs for the third-person singular (*intends*, *depicts*, *pretends*). Although many orthographic strings belong to both classes (*shoots*, *dreams*, *limps*), since these two use cases necessarily have very different meanings, a purely semantic account of morphological family organization predicts that we should be able to separate them.

We constructed our CDV by averaging the vectors of 100 random words (from among the 15,656 words ending in *s*) that were unambiguously NOUN+*s* (i.e., words such as *bonanzas*, *mandolins*, and *waistlines*, for which the root does not easily function as a verb, though English seems to allow its speakers to verb any noun if they try hard enough). Figure 2 shows that distance to this vector is a good predictor of ending with the morpheme *-s*. A total of 35.5% of final-*s* words are in the first decile of the dictionary after sorting by proximity to the CDV [ $\chi^2(1) = 139.24, p = 1.95\text{E-}32$ ], making 71.2% of words in the decile final-*s* words.

We again examined the first 1,000 words manually. Of these, 96.9% were NOUN+*s* words [ $\chi^2(1) = 525.49, p = 1.35\text{E-}116$ ].

We used the same method with the verbs, constructing a CDV by averaging 100 words that were unambiguously

<sup>1</sup> The exact *p* value is too small to be computed using off-the-shelf computational tools (to be precise, it is lower than  $2^{-1022}$ , or  $2.22\text{e-}308$ ). We will henceforth represent such inconveniently low numbers as  $\approx 0$ .

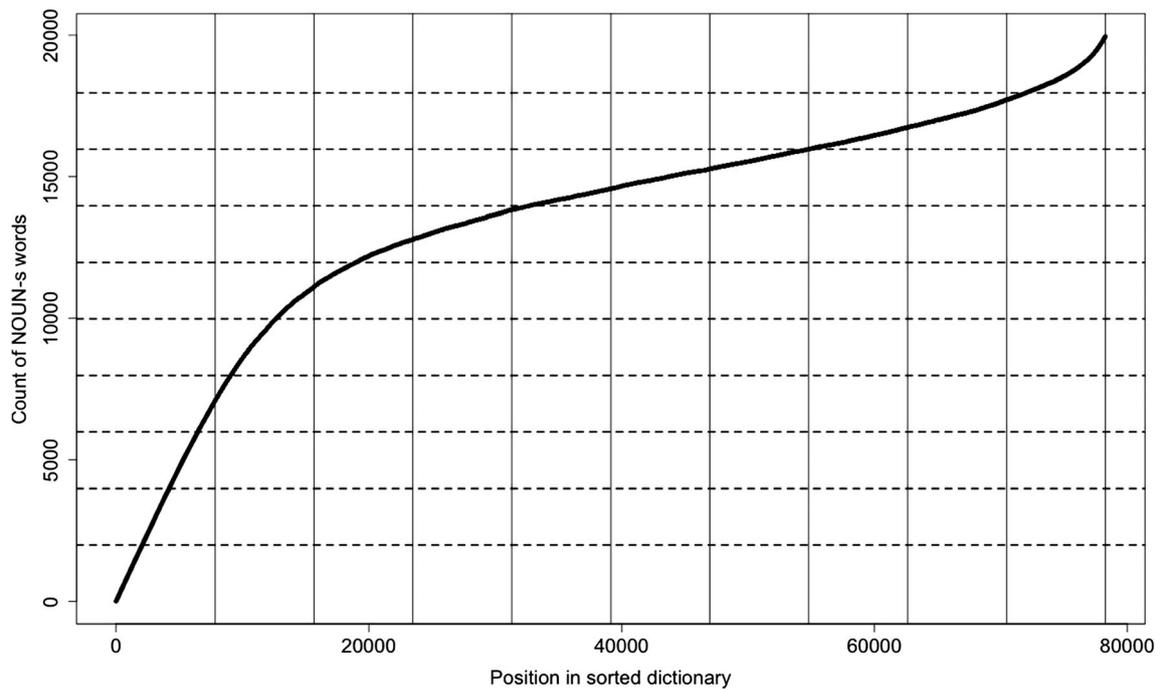
**Table 1** Summary of success in predicting membership in 32 different word categories

Category	N	First decile				First 1,000 words			
		Observed	Expected	Chi-square	p	Observed	Expected	Chi-square	p
<b>Affixation</b>									
in-X	2,046	607	205	248.29	3.06E-56	172	26	113.13	1.01E-26
mis-X	314	254	31	319.47	9.45E-72	192	4	262.14	2.93E-59
noun + s	15,656	2,249	1,566	139.24	1.95E-32	969	200	525.488	1.35E-116
un-X	2,069	1,334	207	1,313.30	7.28E-288	477	26	460.33	2.04E-102
Verb + s	19,959	3,058	1,996	255.51	8.18E-58	1,000	255	456.61	1.32E-101
X-able	906	808	91	1,134.92	4.24E-249	435	12	531.37	7.10E-118
X-ally	685	608	69	848.35	8.29E-187	343	9	426.50	4.69E-95
X-ate	807	502	81	475.93	8.22E-106	212	10	213.12	1.44E-48
X-ed	6,965	5,078	697	5,677.04	~ 0	924	89	742.25	9.69E-164
X-ful	180	165	18	240.17	1.81E-54	84	2	102.73	1.92E-24
x-ible	137	118	14	158.11	1.47E-36	77	2	100.05	7.43E-24
X-ic	1,202	744	120	703.51	2.58E-155	205	15	180.62	1.77E-41
X-ion	1,813	1,305	181	1,440.55	~ 0	493	23	499.13	7.35E-111
X-ish	306	140	31	96.42	4.64E-23	75	4	73.27	5.66E-18
X-ist	538	475	54	659.07	1.19E-145	304	7	398.94	4.69E-89
X-ly	3,208	3,201	321	5,221.09	~ 0	994	41	1,046.28	7.84E-230
X-ness	1,277	1,233	128	1,920.64	~ 0	652	16	820.01	1.21E-180
X-ology	125	120	13	183.94	3.35E-42	103	2	167.50	1.30E-38
X-ous	535	393	54	441.56	2.48E-98	162	7	168.82	6.68E-39
<b>Parts of speech</b>									
Adjectives	9,433	2,943	943	1,296.36	3.50E-284	832	207	397.87	7.99E-89
Adverbs	2,558	2,203	256	2,968.33	~ 0	965	56	1,011.06	3.53E-222
Nouns (Abstract)	2,828	2,828	283	4,626.98	~ 0	949	62	947.59	2.22E-208
Nouns (Concrete)	3,595	2,571	360	2,815.68	~ 0	961	39	987.42	4.88E-217
Prepositions	100	95	10	144.86	1.15E-33	83	2	134.24	2.42E-31
Verbs	13,301	3,985	1,330	1,657.39	~ 0	1,000	170	615.89	2.93E-136

VERB+s, such as *instigates*, *weans*, and *rescinds*. These words are quite rare, in part because so many words can function as both nouns and verbs in English, a fact of which we take advantage in Experiments 3 and 4 below. We had to look through nearly 1,400 randomly selected words ending with *s* to find 100 unambiguous VERB+s exemplars. Figure 3 reflects this rarity. The graph of the number of words ending with *s* as a function of distance from the averaged VERB+s vector has a very different shape from the other graphs we have considered, with a small sharp rise at the beginning, following which words ending in *s* are distributed nearly randomly—that is, with a slope of approximately 1 in the decile/decile graph. In all, 15.5% of final-*s* words are in the first decile of the sorted dictionary [ $\chi^2(1) = 255.51$ ,  $p = 8.18E-58$ ]. However, this low number does not reflect a failure in classification accuracy, but rather reflects the relatively small size of the VERB+s morphological family as compared to the size of

the category of words ending in *s*. Manual inspection confirmed that every one of the 1,000 words closest to the average VERB+s vector is a third-person singular verb [ $\chi^2(1) = 456.6$ ,  $p = 1.32E-101$ ].

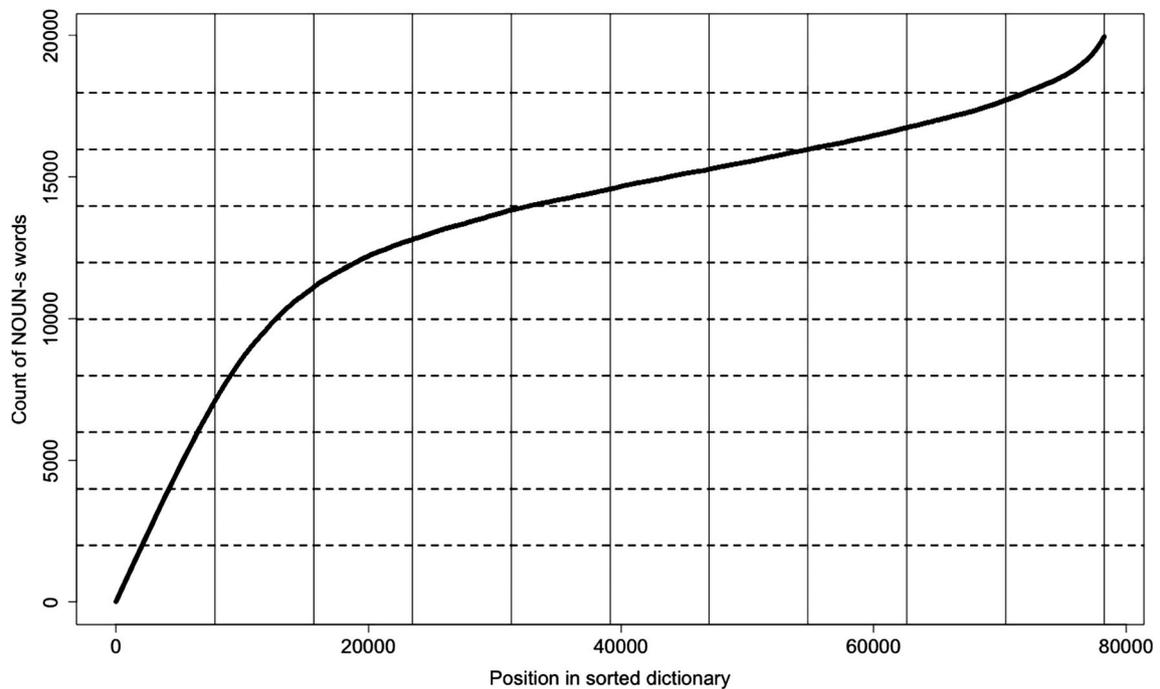
The relationship between the NOUN+s and VERB+s vectors is shown graphically in Fig. 4. Although most standardized PC values in both vectors cluster near zero, suggesting that they probably contribute little to the classification, there are a few very large differences. The two vectors are weakly but reliably negatively correlated, at  $r = -.15$  ( $p = .01$ ). Their respective estimates across the entire dictionary correlate at  $r = -.53$  ( $p \approx 0$ ). These observations suggest that there is a sense in which nouns and verbs (or, at least, a word's degree of "nouniness" and its degree of "verbiness") are each other's opposites in this co-occurrence space. We present behavioral evidence supporting this claim, and define what it means in more precise terms, in Experiments 3 and 4 below.



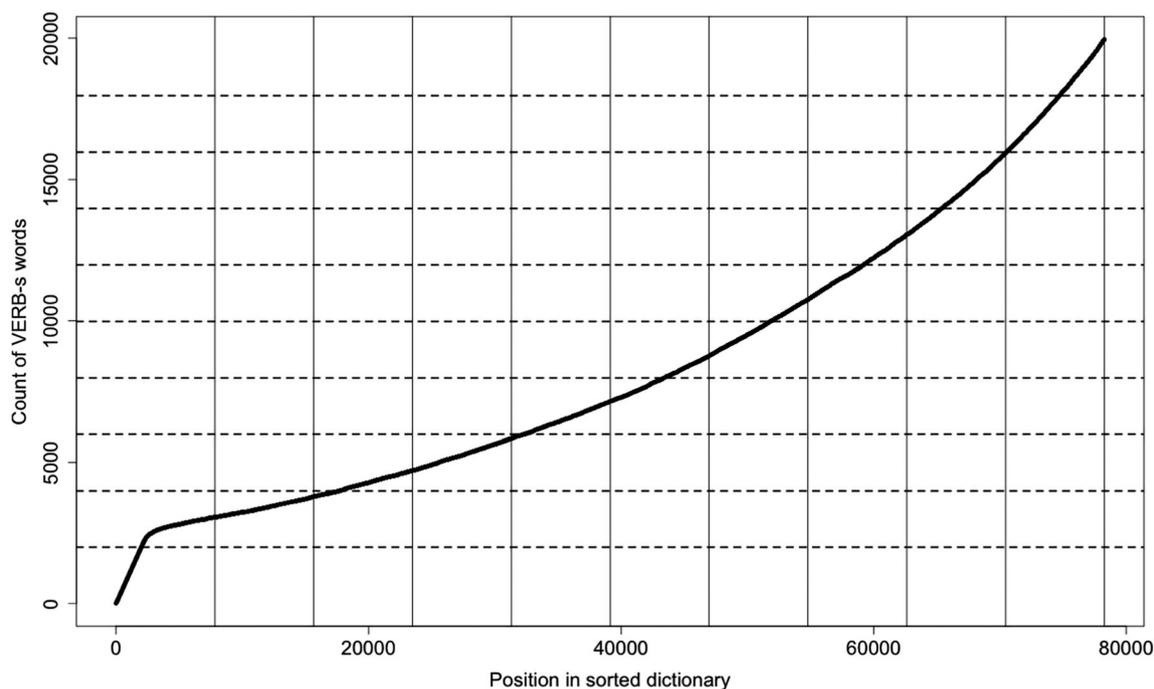
**Fig. 1** Numbers of words ending with *ed* as a function of similarity to the vector defined by averaging 100 random verbs ending with *ed*. The horizontal and vertical lines define deciles among words ending with *ed* and among all words in the dictionary, respectively

**Affix un-X** For our third detailed example, we consider the prefix *un-*. This is an interesting case for two reasons. One is that switching to a prefix means that we will not even have the same part of speech (let alone the same *meaning*) included in the

category of interest, since there are many different classes of words to which this prefix can attach (e.g., verbs such as *untangle*, *untangles*, *untangled*; gerunds such as *undoing*; adverbs such as *unwillingly*; nouns such as *unbelievability*;



**Fig. 2** Numbers of words ending with *s* as a function of similarity to the vector defined by averaging 100 random NOUN+s words. The horizontal and vertical lines define deciles among words ending with *s* and among all words in the dictionary, respectively



**Fig. 3** Numbers of words ending with *s* as a function of similarity to the vector defined by averaging 100 random words ending with VERB+*s*. The horizontal and vertical lines define deciles among words ending with *s* and among all words in the dictionary, respectively

and adjectives such as *unclean*). The second reason we chose *un-* is that there are many semantically related negator affixes in English, which provides us with an opportunity to cross-validate our method by comparing the classification performance of the CDVs from two semantically similar affixes, head to head.

There are 2,069 words in our dictionary that begin with *un*. As is shown in Fig. 5, distance from the CDV defined as the averaged vector of a random 100 of those words is a good predictor of beginning with *un*. A total of 64.4% of such words are in the first decile of the dictionary after sorting it by similarity to the *un* CDV [ $\chi^2(1) = 1,313.30$ ,  $p = 7.28E-288$ ].

As above, we manually examined the 1,000 words closest to that CDV. In all, 477 of those 1,000 words were words prefixed with *un-* [ $\chi^2(1) = 460.33$ ,  $p = 2.04E-102$ ]. This is 22.6% of all words beginning with those two letters. Of the remaining 523 words in the top 1,000 neighbors, 47% (246) were words affixed with one of 11 affixes closely related in meaning to *un-*: *in-* (106 occurrences), *-less* (34 occurrences), *dis-* (30 occurrences), *im-* (25 occurrences), *ir-* (22 occurrences), *mis-* (eight occurrences), *il-* (six occurrences), *non-* (five occurrences), *a-* (four occurrences of this initial letter as a recognizable affix), *under-* (four occurrences), and *anti-* (two occurrences). Many of the remaining words had connotations of absence or emptiness—for example, *abandoned*, *bereft*, *devoid*, *empty*, *futile*, *isolated*, and *purposelessly*.

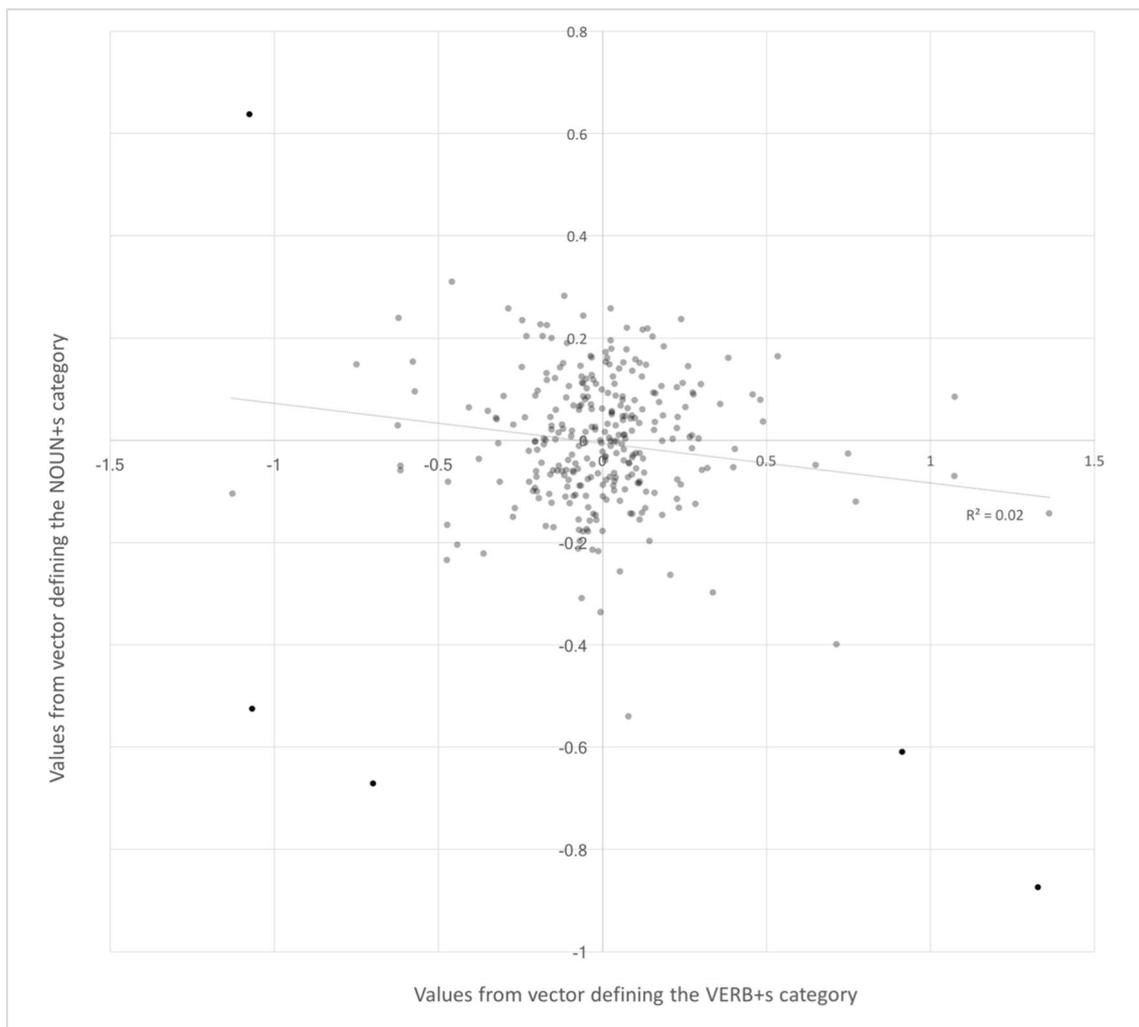
As a validity check on our methods, we modeled the most numerous semantically similar class, *in-*. There are 2,045 words in our dictionary that begin with *in*, although of course not all of those are affixed with the negator morpheme (e.g.,

*index*, *immate*, *invention*). We randomly selected 100 that were prefixed with *in-* as a negator to make our CDV.

The results are summarized at the top of Table 1. A total of 29.7% of all *in-* words were in the first decile of the dictionary after sorting it by proximity to the CDV [ $\chi^2(1) = 248.29$ ,  $p = 3.06E-56$ ]. The closest 1,000 words contained 17.2% of all *in-* words [ $\chi^2(1) = 113.13$ ,  $p = 1.01E-26$ ].

Our main interest was in verifying that both the separately derived CDVs and their estimates were similar for *un-* and *in-*, as one would expect them to be. The 300 vector values were correlated at  $r = .49$  ( $p = 2.8E-19$ ). The category-defining estimates they produced are graphed in Fig. 6. Those estimates are highly correlated,  $r = .82$  ( $p \approx 0$ ), across the 78,278 words in our dictionary.

**Nouns** We have demonstrated that we can use word co-occurrence information by itself to accurately classify words into their morphological families. In doing so, we are really classifying them into semantic categories of their part of speech, as is demonstrated by the morphological “misclassifications”—for example, the inclusion of irregular past tense verbs with verbs selected to end in *-ed*. We now turn to classifying part of speech directly, and also to examining whether the information in the CDVs is sufficient for affixing words. We will discuss verbs, nouns, and prepositions here. Summaries from other part-of-speech classifications are included at the bottom of Table 1. They may be summarized briefly by saying that every part of speech we have considered can be identified with a CDV with very good accuracy.



**Fig. 4** Correlations of normalized vector values for the CDVs in the NOUN+s and VERB+s categories. The five points in black are extreme ( $|z| > 0.5$ ) on both measures

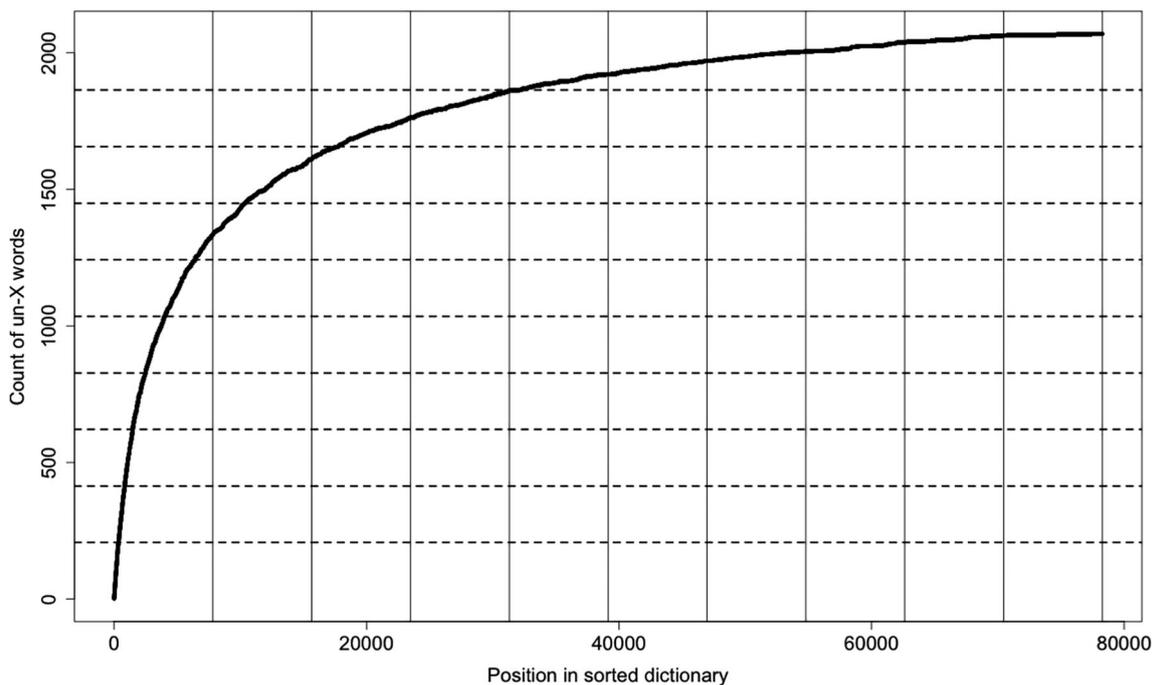
Although nouns are treated as a single word class, semantically they divide naturally into two subclasses. *Concrete* nouns are those that refer to tangible objects or events in the world, and *abstract* nouns are those that do not. Hollis and Westbury (2016) presented evidence suggesting that concreteness is a natural dimension of lexical semantics. We therefore proceed here by treating abstract and concrete nouns as separate categories. In doing so, we will present further evidence of their natural dissociation.

We identified part of speech using the public-domain part-of-speech dictionary from the Moby project (Ward, 1996). We defined concreteness using the norm set from Hollis, Westbury, and Lefsrud (2017), which extrapolated human norms to our entire dictionary. The overlap between these resources consists of 47,426 words. A concrete noun ( $N = 3,072$ ) was defined as a word tagged as a noun that had an estimated concreteness more than 1.5 *SDs* higher than the average estimate over all part-of-speech-tagged words, and an abstract noun ( $N = 2,828$ ) was defined as a noun-tagged

word that had an estimated concreteness lower by at least 1 *SD* than the average estimate. These asymmetrical cutoffs were selected by hand to make the two categories roughly the same size, because a symmetric cutoff of  $\pm 1.5$  *SDs* defined only 22 nouns as abstract.

Using the 47,426-word lexicon, we proceeded in the same way as above, by selecting 100 random words from each category of interest, averaging them together, and computing the similarity of every word's vector to that CDV.

The results for concrete nouns are shown in Fig. 7. Similarity to the average concrete-noun vector was an excellent predictor of a word being a concrete noun. In all, 2,571 (71.5%) of the 3,595 concrete nouns in the dictionary fell in the first decile of the sorted dictionary [ $\chi^2(1) = 2,815.68, p \approx 0$ ], resulting in that decile being composed of 82.6% concrete nouns. In addition, 96.1% of the words in the 1,000 words closest to the CDV were concrete nouns [ $\chi^2(1) = 987.42, p = 4.88E-217$ ]. All of the remaining words were adjectives that were concrete (both in our own judgment and in the Hollis



**Fig. 5** Numbers of words beginning with *un-* as a function of similarity to the vector defined by averaging 100 randomly selected words beginning with *un*. The horizontal and vertical lines mark deciles among words beginning with *un* and among all words in the dictionary, respectively

et al., 2017, estimates), such as *blondish*, *quilted*, *beveled*, *whorled*, and *crinkly*.

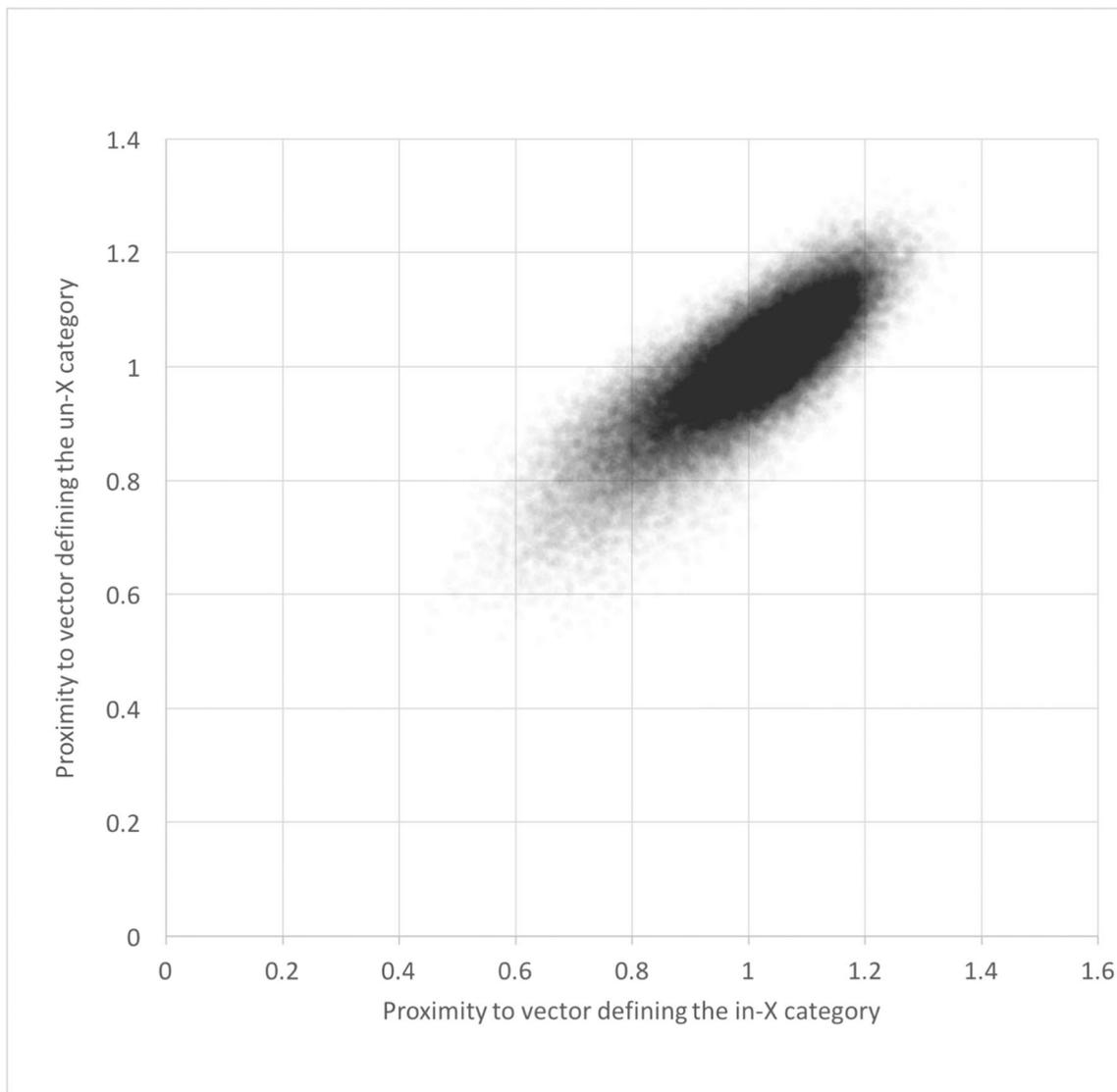
We were also able to compare our distance measure to the Hollis et al. (2017) concreteness estimates that were extrapolated from human judgments. Across the entire dictionary, the negative correlation between distance from the CDV and the extrapolated concreteness judgments is very large,  $r = -.86$  ( $p \approx 0$ )—that is, words closer to the noun-defining vector were estimated as being more concrete.

To fend off any accusations of circularity (since we defined the CDV over exactly the same matrix we used earlier to generate the concreteness estimates), we also correlated the distance to the CDV with the 39,954 human ratings of concreteness from Brysbaert, Warriner, and Kuperman (2014). Among the 28,437 words that appear in both datasets, the correlation of distance with human ratings was  $r = -.74$ ,  $p \approx 0$  (i.e., words closer to the CDV for concrete nouns were rated as being more concrete).

We have now discussed the development of two noun vectors, one plural (the form-defined NOUN+*s* case above) and one singular. Distances from these two CDVs across all words are uncorrelated,  $r = -.08$  ( $p = .18$ ). Since the plural-noun vector is essentially an estimate of the probability that a word is plural, we should be able to use it to find the plural form of any singular noun *X*. We need simply find the word that simultaneously maximizes *X*-ness and plurality probability (which is the geometric equivalent to finding the point that minimizes the distance between two locations in the 300-dimensional space defined by the semantic matrix). As co-occurrence aficionados know, a singular noun's closest

neighbor is often, but not always, its plural. We began with the 1,000 closest words to the CDV for singular nouns. From that pool, we selected the 517 singular count nouns whose plural form was in our dictionary. To compute a given word's plural, we selected the word closest to the NOUN+*s* CDV (i.e., a word judged as plural) that was in the top *N* neighbors of the target word's own vector. We allowed *N* to range from 1 to 7, since we did not know what the optimum value was. As a concrete example, the seven closest neighbors to the word *backpack* are *knapsack*, *duffel bag*, *backpacks*, *rucksack*, *bookbag*, *suitcase*, and *satchel*. When we sort those seven words by increasing distance from the CDV for NOUN+*s* (i.e., by decreasing likelihood of being plural), the closest word is *backpacks*, the correct plural form for the target word *backpack*. In this example, taking only the first- or second-closest neighbors of the word *knapsack* would have missed the correct plural form, since it is the third-closest word.

The results for predicting the plural form of the 517 count nouns are shown in Fig. 8. The closest neighbor of a singular noun was also its plural form 60% of the time. The best hit rate of 71% was achieved using the NOUN+*s* vector, with  $N = 4$ . This was a reliable improvement on classification by just taking the closest neighbor [ $\chi^2(1) = 13.90$ ,  $p = 9.64E-05$ ]. We can compare these results to those reported for a somewhat related task in Mikolov, Yih, and Zweig (2013) and Gladkova, Drozd, and Matsuoka (2016), who used the analogy method described above. Mikolov, Yih, and Zweig estimated both the plural from the singular, as we have done, and the possessive from the singular for 1,000 words, but they reported only combined results. Their method computed the correct word



**Fig. 6** Correlations across 78,278 words between independently derived estimates of membership in the categories of words beginning with *un* and words beginning with *in*. The estimates are correlated at  $r = .82$  ( $p \approx 0$ )

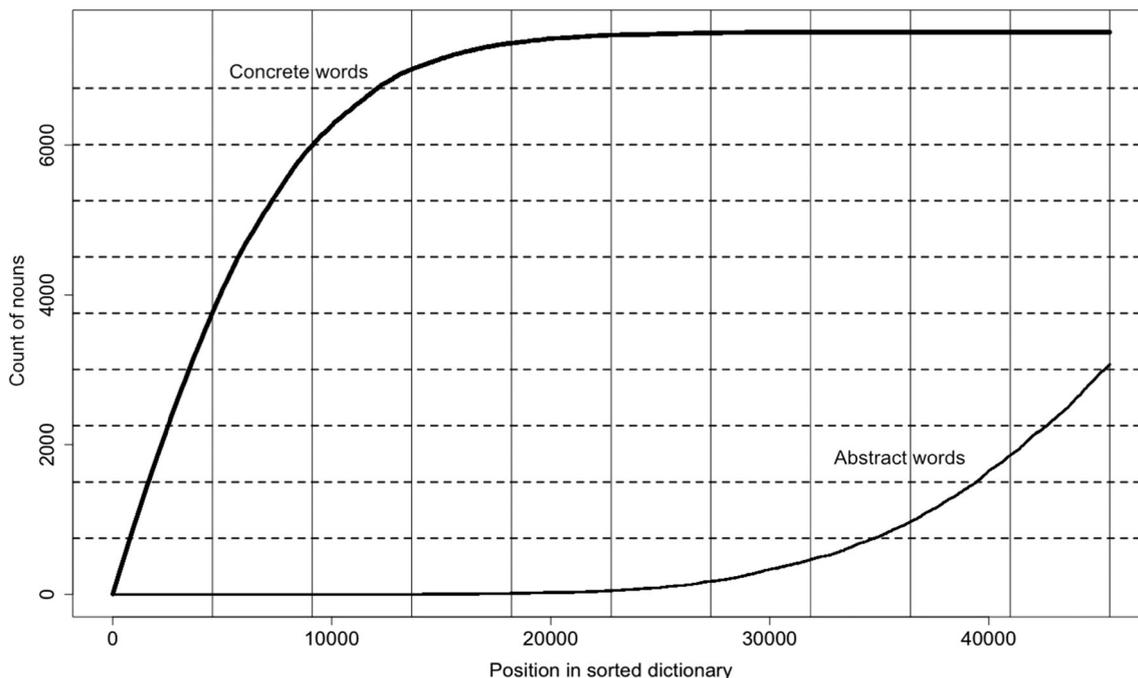
form 29.2% of the time. Gladkova, Drozd, and Matsuoka separated regular and irregular plurals, achieving a hit rate of just over 80% for regular plurals and just over 60% for irregular verbs.

We used the same methods to model the 2,828 nouns that were defined as abstract. We checked the first 1,000 words by hand and confirmed that 94.1% of them were abstract nouns, both by the criterion above and by our own estimate [ $\chi^2(1) = 947.59$ ,  $p = 2.22\text{E-}208$ ]. Of the remaining 59 words, 41 were abstract adjectives (e.g., *numinous*, *inalterable*, *unconfessed*), 13 were abstract plural nouns (e.g., *sophistries*, *antipathies*, *schemata*), and all but one (*inhere*) of the remaining five were abstract adverbs (e.g., *axiomatically*, *paradoxically*, *irreducibly*). The incidence of abstract nouns as a function of proximity to the CDV is shown in Fig. 9. All 2,828 nouns defined as abstract by our criterion fell in the first decile of

the dictionary after it was sorted by proximity to the CDV [ $\chi^2(1) = 4,626.98$ ,  $p \approx 0$ ]. The CDVs for concrete and abstract nouns are strongly negatively correlated ( $r = -.425$ ,  $p = 1.38\text{E-}14$ ), suggesting that abstract and concrete words are encoded very differently.

We again cross-validated our results by correlating the distance to the CDV with the 39,954 human ratings of concreteness from Brysbaert, Warriner, and Kuperman (2014). Among the 28,437 words in both datasets, the correlation of distance with human ratings was  $r = .58$  (i.e., words closer to the CDV for abstract nouns were rated as more abstract,  $p \approx 0$ ).

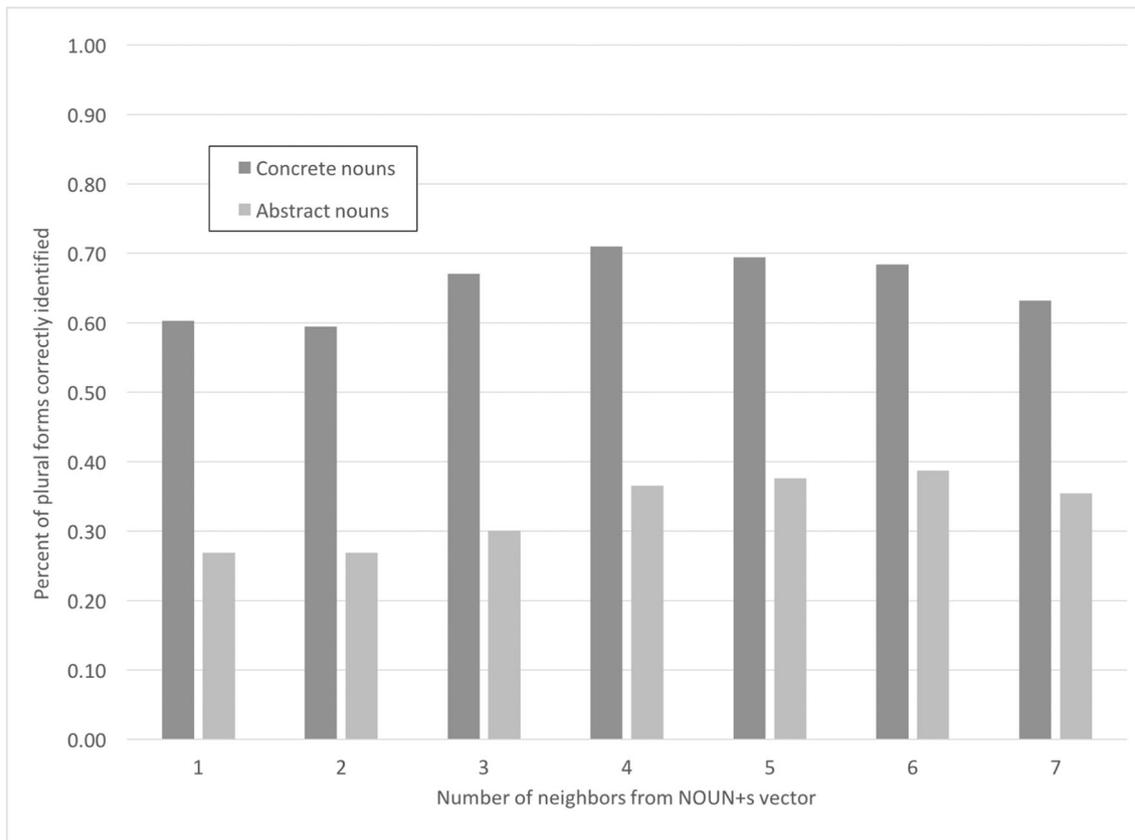
The comparisons of our estimates to the Brysbaert et al. (2014) norms are not quite equivalent, since humans judge the whole range at once (i.e., estimating the relative weightings of concreteness vs. abstractness), whereas we estimated concreteness independently for each pole,



**Fig. 7** Numbers of concrete and abstract nouns as a function of similarity to the vector defined by averaging 100 randomly selected concrete nouns. The horizontal and vertical lines mark deciles among nouns and among all words in the dictionary, respectively

producing correlations of very different magnitude (i.e.,  $r = .58$  for abstractness vs.  $r = -.74$  for concreteness). We

combined the estimated values of concreteness and abstractness using linear regression over the 28,437 words. The best



**Fig. 8** Percentages of plural concrete nouns ( $N = 517$ ) and abstract nouns ( $N = 93$ ) correctly identified, as a function of number of a singular word’s neighbors examined on the NOUN+s vector

model produced estimates that correlated with human judgments at  $r = .757$  ( $p \approx 0$ ), with beta weights of  $-4.58$  on the concreteness estimates and  $1.38$  on the abstractness estimates, and an intercept of  $6.26$ . Although the improvement was small, by Fisher's  $r$ -to- $z$  test, this combined model was still better at predicting the human judgments than was using only the distance from the CDV for concrete nouns ( $z = 3.87$ ,  $p = .0001$ , one-tailed).

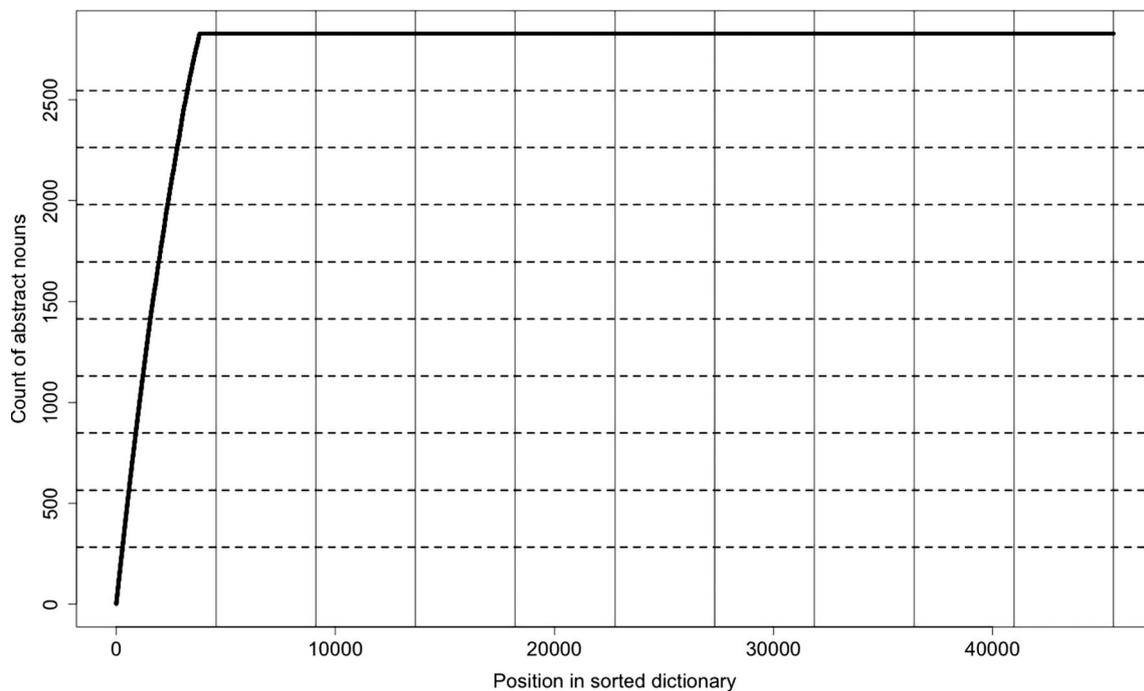
Of the 949 abstract nouns in the 1,000 words closest to their CDV, only 93 had a plural form in the matrix. Most abstract nouns are treated as mass nouns and are not easily subject to pluralization. We use the same method as above to calculate the plural form of the abstract nouns, using the NOUN+ $s$  vector (Fig. 8). The closest neighbor to a singular abstract noun was its plural form just 26.8% of the time. Sorting the  $N$  words closest to the CDV for NOUN+ $s$  achieved a maximum hit rate of 38.7% when  $N = 6$ , which was a marginally reliable improvement on just choosing the closest neighbor of each singular word [ $\chi^2(1) = 2.95$ ,  $p = .04$ ], and also an improvement on the results for the similar tasks (which did not discriminate between abstract and concrete nouns) reported in Mikolov, Yih, and Zweig (2013).

**Verbs** There are 13,301 words marked as verbs in the dictionary. We randomly chose 100 of them to create a CDV for the word class. The distribution of marked verbs as a function of proximity to that vector is shown in Fig. 10. In all, 30% of all verbs were in the first decile of the dictionary [ $\chi^2(1) = 1,657.4$ ,  $p \approx 0$ ]. As with plural nouns, this number is low because the

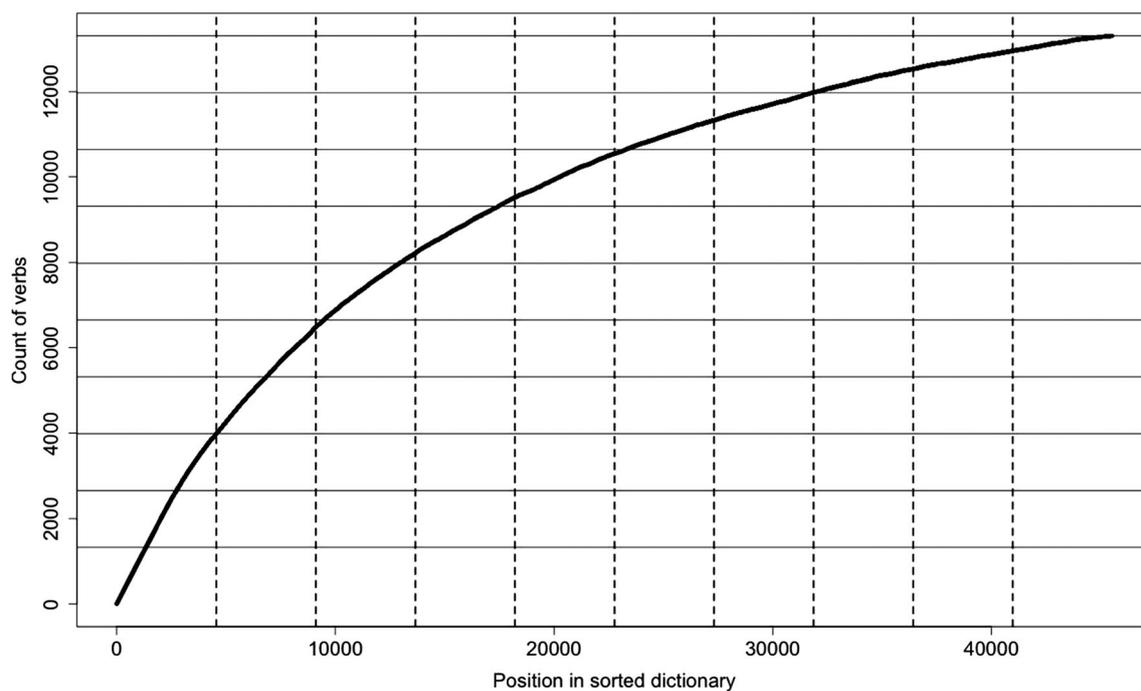
class is large. Having 30% of 13,301 verbs in the first 4,553 words of the sorted part-of-speech-marked dictionary means that 87.6% of the words in that decile were present tense verbs. Every one of the first 1,000 words was a present tense verb [ $\chi^2(1) = 615.89$ ,  $p = 2.93E-136$ ].

We used exactly the same method that we used above to pluralize nouns to compute the past tense of those 1,000 verbs, by sorting the  $N$  closest neighbors of each word by their proximity to the CDV for VERB+ $ed$ , with  $N = 1$  to 7. The best performance (a tie with  $N = 6$  or 7) was 78.6%. Simply taking the first neighbor of each word would have scored just 22%, suggesting that our method was much better [ $\chi^2(1) = 1,212.07$ ,  $p = 7.25E-266$ ]. Mikolov, Yih, and Zweig (2013) also tried to compute the past tense from a verb in an analogy task, as described above, but again combined the results with those from other tasks (predicting the third-person singular present from either the base or past tense form). They reported a combined success rate of 62.2%. In estimating past tense forms from the third-person singular form, Gladkova, Drozd, and Matsuoka (2016) reported a success rate (estimated from their Fig. 1, since it was not reported explicitly) of just under 50%.

We performed the same calculation for computing the present participles of the 1,000 verbs, using proximity to the CDV for VERB+ $ing$ . The present participle was the first neighbor of the word 45.9% of the time. The best performance attained using proximity to the CDV for VERB+ $ing$  was 80.8%, when  $N = 7$ . By comparison, using the offset method, Gladkova, Drozd, and Matsuoka (2016) reported a success rate of about



**Fig. 9** Numbers of abstract nouns as a function of similarity to the vector defined by averaging 100 randomly selected abstract nouns. The horizontal and vertical lines mark deciles among abstract nouns and among all words in the dictionary, respectively



**Fig. 10** Numbers of verbs as a function of similarity to the vector defined by averaging 100 randomly selected verbs. The horizontal and vertical lines mark deciles among abstract nouns and among all words in the dictionary, respectively

70% in predicting the present participle from the infinitive form.

Table 1 presents results from many other categories that we do not consider in detail in the text. These results are represented in summary form in Fig. 11, which shows the largest-magnitude correlations between the estimates for all the word classes in Table 1. There are many large negative and positive relations, all of which are semantically predictable (i.e., strong correlations between adjectives and *X-ous* words; strong correlations between *X-ible* and *X-able* words; strong correlations between adverbs and adjectives; few or no relations between part-of-speech categories). Also notable is that the distance to the CDV for abstract nouns is strongly correlated with distance to the CDVs of other word types (notably, adjectives and adverbs), whereas distance to the CDV for concrete nouns shows no such relations. The ability of a VERB+*ed* CDV to identify irregular past tense forms and the strong correlation between the CDVs of semantically related affixes (such as *in-* and *un-* or *X-ible* and *X-able*) suggests that the CDV defined by averaging morphologically similar words is not a “representation of the affix,” but rather a representation of the *semantics* of the affix.

## Part 2: Empirical validation

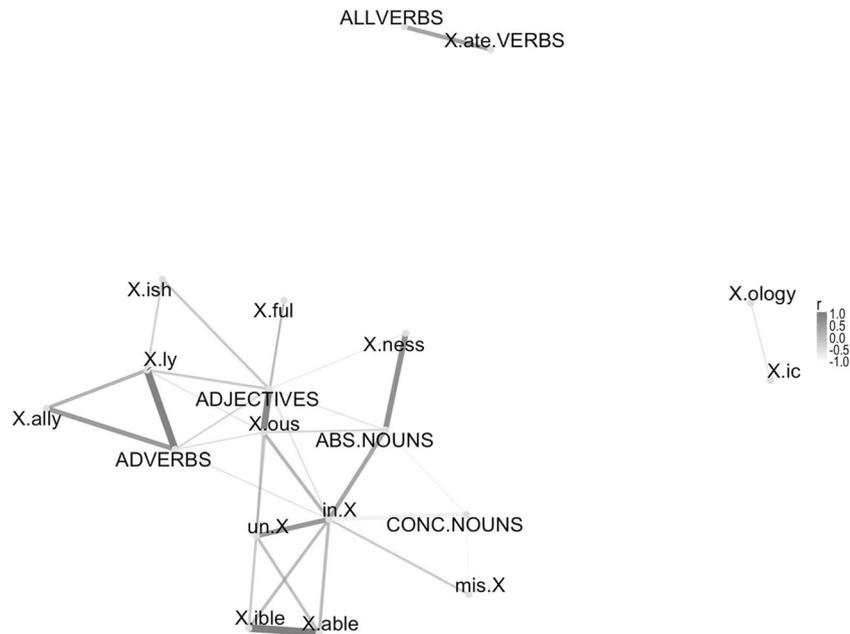
### Cross-validation

To cross-validate these part-of-speech classifications, we selected 500 random words from each of the five categories of abstract and concrete nouns, verbs, adjectives, and adverbs. We

classified these simply by measuring distance from each of the CDVs, classifying the word in the category to whose CDV it was closest. The results are shown in Table 2. The 500 randomly selected verbs and concrete nouns were both categorized without any errors [ $\chi^2 = 666.67, p \approx 0$ ]. Abstract nouns were categorized with 97.8% accuracy [ $\chi^2(4) = 627.589, p < 2E-16$ ], and adverbs with 88% accuracy [ $\chi^2(4) = 465.54, p < 2E-16$ ]. The worst category was adjectives, which were categorized with 48.8% accuracy [ $\chi^2(4) = 124.37, p < 2E-16$ ]. In all, 14% of the errors in categorizing adjectives were due to miscategorizing abstract adjectives (e.g., *impious*, *bibliographical*, *vicarious*, *unmerited*) as abstract nouns, and 20% were due to miscategorizing adjectives as concrete nouns. These errors were sometimes due to categorical ambiguity (e.g., adjectives such as *fleet*, *teen*, *facial*, *blind*, and *Brazilian* can also be used as nouns). The part-of-speech dictionary again seems unreliable on some relevant words—for example, classifying many adjectives (e.g., *icebound*, *flammable*, *gateless*, and *rhomboid*) as concrete nouns and classifying many nouns (e.g., *transmutation*, *convalescence*, *halt*, and *jingoism*) as adjectives. After removing words that were either categorically ambiguous or incorrectly classified in the part-of-speech dictionary as adjectives, the correct classification rate for the remaining 442 adjectives was 55%.

### Behavioral measures

The demonstration that it is possible to identify part-of-speech membership from a skip-gram matrix does not mean that that the human mind is sensitive to the relationship between co-



**Fig. 11** Largest-magnitude correlations ( $|r| > .6, p \approx 0$ ) between distance to the CDVs for all categories in Table 1. Note that the threshold used to make this figure comprehensible is extremely high: With 78,278 words, a

correlation of just  $|r| = .05$  would be reliable at  $p < 2e-44$ . The distance between unconnected clusters is arbitrary and has no interpretation

occurrence and word class, but it does suggest a straightforward experimental test of that claim. Since part-of-speech membership defined by distance from a CDV is a continuous measure, we can hypothesize that people should be quicker to make word category decisions for words that are closer to their CDV. In this section we report the results of four experiments showing that this is true.

Testing the hypothesis depends on one assumption that is implicit above, which is that there is a single canonical CDV for each word category. We defined our CDVs by selecting words randomly and showed that those CDVs were good at classifying words into parts of speech. Since the words in each CDV were random, the implicit assumption was that it didn't matter which words we chose, because the CDV would be roughly the same for any set of words.

To explicitly test the claims that there is a canonical CDV for each word class and that humans are sensitive across the range of distance from such a CDV, we here focus on the two classes of words: plural nouns and third-person verbs.

Although this choice was arbitrary, we opted for third-person agreement because this allows for not only verbs that have an active human subject (e.g., *he*, *she*) but also such verbs as *percolates* and *mineralizes*, which obligatorily take a nonhuman subject, except perhaps in fictional settings. We chose regular plural nouns because in their standard form, these words look very similar to third-person verbs (with a final *-s*), so there were not likely to be obvious formal clues to class membership. We also took advantage (in Exps. 3 and 4) of the fact that many words that are plural nouns can also be used as verbs.

To validate the CDV for each of the two classes of interest, we defined four CDVs for each class. Each one was defined by averaging together 500 randomly selected disjunct exemplars from the category. Within each word class, we correlated both the vector values for each of the four independently defined CDVs and the distances from that vector for every word in a dictionary of 78,278 words. The results, shown in Table 3, can easily be summarized: All vectors (and therefore the

**Table 2** Correct classification proportions for classifying 500 words per part-of-speech category according to the CDV to that each word was closest

	Actual category				
	Nouns (Concrete)	Nouns (Abstract)	Verbs	Adverbs	Adjectives
Classified as Nouns (Concrete)	<b>1</b>	0	0	0	0
Classified as Nouns (Abstract)	.002	<b>.978</b>	0	0	.02
Classified as Verb	0	0	<b>1</b>	0	0
Classified as Adverb	.026	.03	.032	<b>.88</b>	.032
Classified as Adjective	.204	.14	.124	.044	<b>.488</b>

distances of words from those vectors) were essentially identical,  $r \geq .99$ . This strongly suggests that, at least for these two classes, there is a single canonical CDV. We nevertheless averaged all four CDVs in each word class in order to define our final CDV for each class.

## Experiment 1: Noun decision

### Method

**Participants** The participants were 63 people (46 females, 17 males) who reported themselves to be right-handed native English speakers. They had an average [*SD*] age of 18.8 [1.3] years, and an average [*SD*] number of years of education since beginning grade school of 13.4 [0.69] years. They participated in this experiment in return for partial course credit in an introductory psychology class at the University of Alberta.

**Materials** To select stimuli for the experiment, we sorted our dictionary of 78,278 words by decreasing proximity to each of the CDVs. We then selected 500 verbs by taking every fifth consecutive word, unless that word was either irregular, not a member of the class, or a member of both classes (i.e., a word that could serve as both a verb and a noun), in which case we took the word that was next closest to the CDV and also acceptable. The cosine distance of the verbs to the verb CDV ranged from .34 to .91. We similarly selected 500 nouns by taking approximately every tenth consecutive word (because there are many more nouns than verbs) in order to select 500 “pure” nouns whose cosine distance covered about the same range, from .52 to .92. We call the distance from a category’s respective CDV its *CDV-distance*.

Each participant was asked to make 300 decisions, as described in the next section. The 300 stimuli (150 verbs and 150 nouns) were selected randomly (without replacement until all had been used) individually for every participant.

**Procedure** Data were collected using three Apple G4 Macintosh Minis with 17.1-in. monitors. The screens’ resolutions were set to 1,280 × 1,024. Words were presented one at a time, in 90-point black Times font, in the same location centered on a white screen. Each word was preceded by a cross bar (“+”) that remained on screen for a uniformly sampled random time between 500 and 1,500 ms. The interstimulus interval was 1,000 ms.

The task was a go/no-go noun decision task. Participants were shown written instructions, which were also read out loud by a research assistant. The instructions explained that they would see one word at a time on the screen and asked them to press the space bar as quickly as possible, without

sacrificing accuracy, if they recognized the word as a noun, defined in the instructions as “a word that names an animal, place, thing, or idea.” In the absence of a response, the word disappeared after 2,000 ms, ending the trial.

To accustom participants to the task, each one saw the same four trials to begin with, consisting of the words *plucks*, *gets*, *electrons*, and *pinballs*, presented in randomized order. These four trials were discarded prior to analysis.

### Results

We began by computing the percentage of trials that were answered correctly by each participant. Three participants were eliminated because they scored at or significantly below ( $p < .05$ ) chance levels. The average [*SD*] percent correct of the remaining participants was 83.7% [7.2%]. We eliminated three more participants who scored more than 2 *SDs* below this average. The remaining participants had an average [*SD*] score of 84.8% [5.7%].

After matching our stimuli to measures of length, logged frequency (from Shaoul & Westbury, 2006), orthographic neighborhood size (also from Shaoul & Westbury, 2006), and estimates of valence, arousal, dominance, and concreteness (from Hollis et al., 2016), we found that 18 of the nouns did not have complete information. We eliminated these stimuli and analyzed the correct reaction times (RTs) for the 482 remaining nouns. These nouns had been correctly responded to an average [*SD*] of 13.7 [4.5] times each. The variance is attributable to the fact that some nouns (e.g., *abutments*, *ingrates*, *ascetics*, *rubes*) were poorly recognized.

We analyzed the correct accuracy data for plural nouns only using binomial linear mixed-effect (LME) modeling, with predictors scaled before they were entered. We measured

**Table 3** Correlations between vectors and distances of all words for four disjunct CDVs defined for plural nouns and third-person verbs

	List 1	List 2	List 3	List 4
<b>Nouns</b>				
List 1	(1)	.991	.988	.989
List 2	.978	(1)	.992	.993
List 3	.975	.981	(1)	.993
List 4	.978	.981	.985	(1)
<b>Verbs</b>				
List 1	(1)	.994	.997	.995
List 2	.990	(1)	.997	.996
List 3	.993	.993	(1)	.997
List 4	.990	.991	.993	(1)

The lower triangle of each table shows the correlation between the 300 vector values in each CDV. The upper triangle of each table shows the correlation between the distances to the vectors of 78,278 words.

the fit using the Akaike information criterion (AIC; Akaike, 1974), accepting models that were at least 10 times more likely to minimize information loss than the previous best model. The model analysis and final fixed effects are summarized in Table 4. As is shown in Fig. 12, CDV-distance was a strong predictor of accuracy. Participants made more errors about plural nouns as the vectors for those nouns were further from the plural-noun CDV. The negative beta weight of  $-0.51$  per *SD* is nearly as large in magnitude as the positive weight on LogFreq (0.53), which was the strongest predictor of accuracy among those considered.

We analyzed the correct RT data using LME modeling, with predictors again scaled before they were entered. The best model before entering CDV-distance (AIC = 93,809) included random effects of stimulus and participant, with fixed effects for logged word frequency, length, and concreteness. Adding CDV-distance improved the model substantially (AIC = 93,778), with increasing distance from the plural-noun CDV being associated with longer RTs. The model analysis and final fixed effects are summarized in Table 5. The relationship between the estimated RT and CDV-distance is shown in Fig. 13.

## Discussion

The results of this experiment show clear effects of CDV-distance, after first entering several well-known predictors of RT, on both accuracy and decision time and in the direction hypothesized: Nouns closer to the CDV were more likely to be correctly accepted as nouns, and to be accepted more quickly, than nouns farther from the CDV. The beta weight suggests that the RT effect is approximately equal to the effect of concreteness (though weighted oppositely), about 35 ms per *SD*, or, *pari passu*, about 140 ms across the full range from which our words were drawn.

## Experiment 2: Verb decision

### Method

**Participants** The participants in the second experiment were 61 people (39 females, 22 males) who reported themselves to be right-handed native English speakers and had not participated in Experiment 1. They had an average [*SD*] age of 19.5 [3.0] years, and an average [*SD*] number of years of education since beginning grade school of 13.7 [1.00] years. They again participated in return for partial course credit.

**Materials and procedure** We used the same lists of 500 nouns and 500 verbs that we had used in Experiment 1. The procedure was also identical, except that instead of being asked to make a *go* response to nouns, participants were asked to respond only to verbs, which were defined in the instructions as “a word that names an action or a state of being.”

### Results

We began by computing the percentage of trials that were answered correctly by each participant. One participant was eliminated because they scored at chance level (47.3% correct,  $p = .7$  by exact binomial probability). The average [*SD*] percent correct of the remaining participants was 83.7% [9.1%]. No participant scored more than 2 *SDs* below this average.

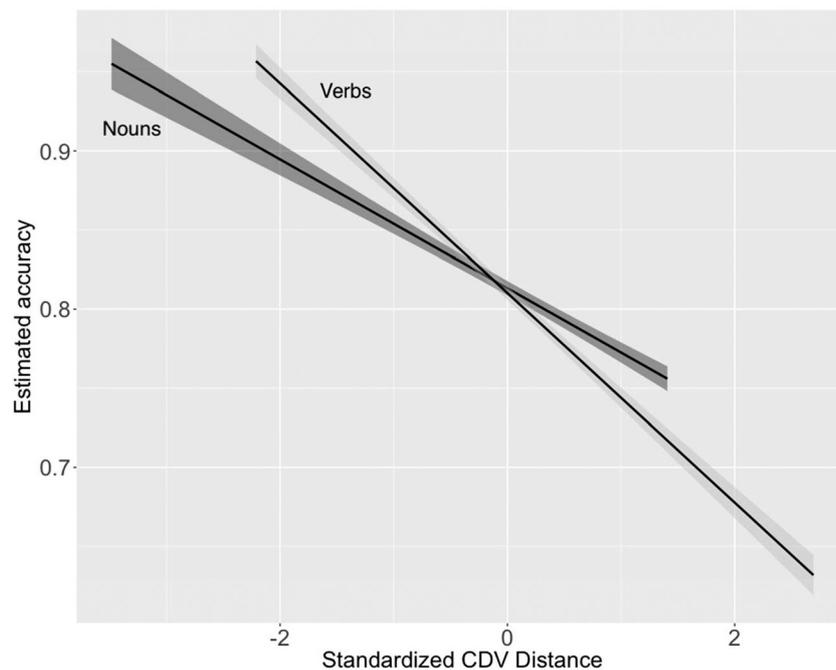
After matching our stimuli to measures of length, logged frequency, orthographic neighborhood size, and estimates of valence, arousal, dominance, and concreteness, from the same sources as referenced above, we found that eight of the verbs did not have complete information. We eliminated these stimuli and analyzed the correct RTs for the remaining 492 verbs.

As in Experiment 1, we analyzed the correct accuracy data using binomial LME modeling, with predictors scaled before

**Table 4** Model analysis (left) and final fixed effects (right) for LME model predicting accuracy in the “pure noun” judgment task (Exp. 1), with normalized predictors

Model	Specification	AIC	Accept?	Fixed effects	Estimate	<i>SE</i>	<i>z</i>	<i>p</i>
<b>M0</b>	<b>(1   STIM)</b>	<b>7,474</b>	<b>[Base]</b>	(Intercept)	2.35	0.17	14.19	<2E-16
<b>M1</b>	<b>M0 + (1   ID)</b>	<b>6,231</b>	<b>Yes</b>	LogFreq	0.53	0.08	6.39	1.62E-10
M2	M1 + (1   ORDER)	[No convergence]	No	<b>CDV– Distance</b>	<b>– 0.51</b>	<b>0.09</b>	<b>– 5.87</b>	<b>4.29E-09</b>
<b>M3</b>	<b>M1 + LogFreq</b>	<b>6,206</b>	<b>Yes</b>	Length	0.29	0.08	3.6	0.00032
<b>M4</b>	<b>M3 + Length</b>	<b>6,197</b>	<b>Yes</b>	Valence	0.18	0.08	2.19	0.028
M5	M4 + ON	[No entry]	No					
<b>M6</b>	<b>M4 + Valence</b>	<b>6,189</b>	<b>Yes</b>					
M7	M6 + Arousal	[No entry]	No					
M8	M6 + Dominance	[No entry]	No					
M9	M6 + Concreteness	[No convergence]	No					
<b>M11</b>	<b>M6 + CDV-Distance</b>	<b>6,155</b>	<b>Yes</b>					

Accepted model steps are shown in bold.



**Fig. 12** Estimated accuracies at noun and verb judgment (Y-axis) as a function of the standardized distance from each category's CDV (X-axis), with 95% confidence intervals

they were entered. The model analysis and final fixed effects are summarized in Table 6. As is shown in Fig. 12, CDV-distance was a strong predictor of accuracy, as it had been for nouns, with participants again making more errors as the distance from the CDV increased. Verbs that were poorly recognized include *metes*, *flouts*, *wearies*, and *doles*.

We again analyzed the correct RTs using LME modeling. The model analysis and final fixed effects are summarized in Table 7. The best model before entering CDV-distance (AIC = 96,923) included random effects of stimulus and participant, with fixed effects for logged word frequency, orthographic neighborhood size, arousal, and concreteness. Adding CDV-distance improved the model substantially (AIC = 96,886). The beta weight on CDV-distance of 29.6 ms per *SD* is larger in magnitude than the beta weight on concreteness (− 20.7 ms per *SD*), and about half that of frequency, the strongest predictor of RTs among those considered (− 61.4 ms per *SD*). The relationship between estimated RT and CDV-distance is shown in Fig. 13.

## Discussion

The results for verb decisions replicate those for noun decisions, showing strong effects of CDV-distance on both accuracy and decision time in the direction hypothesized: The closer a verb was to the verb CDV, the more accurately and quickly it was recognized as a verb. The effects of CDV-distance were roughly similar for both nouns (36 ms per *SD*) and verbs (29 ms per *SD*, or approximately 145 ms across the full range from which our words were drawn).

## Experiments 3 and 4: Ambiguous noun/verb decisions

In Experiments 1 and 2, we chose target words that were unambiguously members of one word class or the other. However, as we noted in the introduction above, many English words are acceptable as both nouns and verbs (e.g., *pets*, *shelters*, *cans*, *ships*, *faces*). In Experiments 3 and 4, we selected target words that could belong to either word class and had participants make decisions about those words as nouns (against a “pure verb” background; Exp. 3) or as verbs (against a “pure noun” background; Exp. 4). Because these two experiments used exactly the same target stimuli and varied only in the instructions and distractors, we consider them here together.

## Method

**Participants** The participants participated in return for partial course credit. The participants in Experiment 3 (noun decision) were 29 people (eight females, 21 males<sup>2</sup>) who reported themselves to be right-handed native English speakers. They had an average [*SD*] age of 20.1 [3.4] years, and an average [*SD*] number of years of education since beginning grade

<sup>2</sup> As an interesting sociological note, we believe that the gender disparity between Experiments 1 and 2 (mostly female, reflecting our program's enrolment gender disparity) and Experiments 3 and 4 (mostly male) is due to the fact that Experiments 3 and 4 were run in the last few days of the semester. It appears that males are much more likely than females to leave their participation credits until the last minute.

**Table 5** Model analysis (left) and final fixed effects (right) for LME model predicting RT in the “pure noun” judgment task (Exp. 1), with normalized predictors

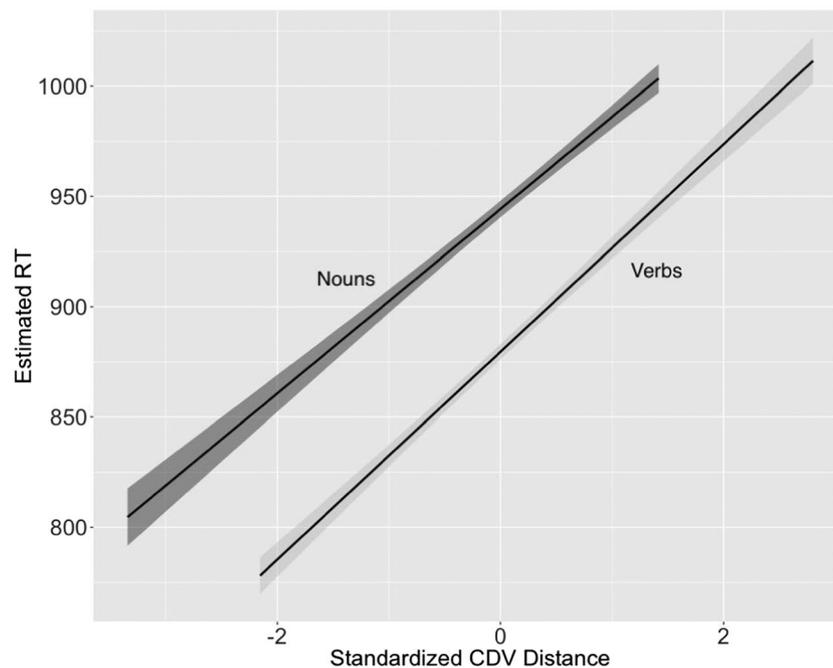
Model	Specification	AIC	Accept?	Fixed effects	Estimate	SE	<i>t</i>
<b>M0</b>	<b>(1   STIM)</b>	<b>94,856</b>	<b>[Base]</b>	(Intercept)	969.15	18.35	52.82
<b>M1</b>	<b>M0 + (1   ID)</b>	<b>93,917</b>	<b>Yes</b>	LogFreq	− 57.4	6.61	− 8.69
M2	M1 + (1   ORDER)	93,919	No	<b>CDV-Distance</b>	<b>35.81</b>	<b>6.8</b>	<b>5.27</b>
<b>M3</b>	<b>M1 + LogFreq</b>	<b>93,874</b>	<b>Yes</b>	Concreteness	− 35.31	6.8	− 5.20
<b>M4</b>	<b>M3 + Length</b>	<b>93,866</b>	<b>Yes</b>	Length	− 18.21	6.34	− 2.87
<b>M5</b>	<b>M4 + ON</b>	<b>93,860</b>	<b>Yes</b>				
M6	M5 + Valence	[ON removed]	No				
<b>M7</b>	<b>M4 + Valence</b>	<b>93,855</b>	<b>Yes</b>				
M8	M7 + Arousal	93,851	No				
M9	M7 + Dominance	[No entry]	No				
M10	M7 + Concreteness	[Valence removed]	No				
<b>M11</b>	<b>M4 + Concreteness</b>	<b>93,809</b>	<b>Yes</b>				
<b>M12</b>	<b>M11 + CDV-Distance</b>	<b>93,778</b>	<b>Yes</b>				

Accepted model steps are shown in bold.

school of 13.8 [0.7] years. The participants in Experiment 4 (verb decision) were 31 people (ten females, 21 males) who reported themselves to be English speakers. Five reported themselves to be left-handed. They had an average [*SD*] age of 19.6 [1.8] years, and an average [*SD*] number of years of education since beginning grade school of 13.7 [0.9] years.

**Materials** To select stimuli for the experiment, we sorted our dictionary by the difference between distance to the noun

CDV and distance to the verb CDV. We then selected the first 320 words that could serve as both a verb and a noun. This number was arbitrary; after 320 it became hard to find more examples, since the words were increasingly distant from both CDVs. The cosine distances from the noun CDV (Noun-CDV) ranged from 0.38 to 1.17. The cosine distances from the verb CDV (Verb-CDV) ranged from 0.52 to 1.25. The difference between the cosine distances (CDV-Diff, defined as Noun-CDV − Verb-CDV) ranged from − 0.64 to 0.80. Examples of words that were much closer to the noun CDV



**Fig. 13** Estimated RTs for correct noun and verb judgments (Y-axis) as a function of the standardized distance from each category’s CDV (X-axis), with 95% confidence intervals

**Table 6** Model analysis (left) and final fixed effects (right) for LME model predicting accuracy in the “pure verb” judgment task (Exp. 2), with normalized predictors

Model	Specification	AIC	Accept?	Fixed effects	Estimate	SE	z	p
<b>M0</b>	<b>(1   STIM)</b>	<b>7,931</b>	<b>[Base]</b>	(Intercept)	2.17	0.18	11.81	<2E-16
<b>M1</b>	<b>M0 + (1   ID)</b>	<b>6,681</b>	<b>Yes</b>	<b>CDV-Distance</b>	<b>- 0.64</b>	<b>0.07</b>	<b>- 9.82</b>	<b>&lt;2E-16</b>
M2	M1 + (1   ORDER)	6,683	No	Length	0.37	0.06	5.81	6.10E-09
M3	M1 + LogFreq	[No convergence]	No	Valence	0.18	0.06	2.76	0.0058
<b>M4</b>	<b>M1 + Length</b>	<b>6,663</b>	<b>Yes</b>					
M5	M4 + ON	[No entry]						
<b>M6</b>	<b>M4 + Valence</b>	<b>6,655</b>	<b>Yes</b>					
M7	M6 + Arousal	[No entry]						
M8	M6 + Dominance	6,651	No					
M9	M6 + Concreteness	[No entry]	No					
<b>M10</b>	<b>M6 + CDV-Distance</b>	<b>6,573</b>	<b>Yes</b>					

Accepted model steps are shown in bold.

than to the verb CDV are *canoes*, *homes*, and *bins*. Example of words that are much closer to the verb CDV than to the noun CDV are *shines*, *taints*, and *interrupts*. Words about equidistant from both CDVs include *brokers*, *bursts*, and *nods*.

The category-ambiguous target words had to be judged against an unambiguous background. We used the 500 pure nouns or pure verbs from Experiments 1 and 2.

Each participant was asked to make 300 decisions. The 300 stimuli (150 ambiguous target words and 150 background words of the appropriate class for the experiment) were randomly selected, without replacement until all had been used, for every participant.

**Procedure** The procedure was identical to that of Experiments 1 and 2, with the addition that, to avoid potential confusion about how to respond to the ambiguous words, we told participants in advance that the words would belong to both the

noun and verb categories, but that they need only respond on the basis of membership in the target category.

## Results

Two participants were eliminated from Experiment 3 because they scored significantly below ( $p < .05$ ) chance levels. We eliminated one more participant in each of Experiments 3 and 4 because they scored more than 2 *SDs* below the average score in their experiment. The remaining participants had average [*SD*] correct decision scores of (Exp. 3) 77.5% [6.4%] and (Exp. 4) 80.2% [6.8%].

We again matched our target stimuli to measures of length, logged frequency, orthographic neighborhood size, valence, arousal, dominance, and concreteness.

The models for predicting accuracy are summarized in Table 8 (noun decision) and Table 9 (verb decision). Both

**Table 7** Model analysis (left) and final fixed effects (right) for LME model predicting RT in the “pure verb” judgment task (Exp. 2), with normalized predictors

Model	Specification	AIC	Accept?	Fixed effects	Estimate	SE	t
<b>M0</b>	<b>(1   STIM)</b>	<b>98,450</b>	<b>[Base]</b>	(Intercept)	896.23	18.7	47.93
<b>M1</b>	<b>M0 + (1   ID)</b>	<b>97,128</b>	<b>Yes</b>	LogFreq	- 61.46	4.99	- 12.31
M2	M1 + (1   ORDER)	97130	No	<b>CDV-Distance</b>	<b>29.64</b>	<b>4.98</b>	<b>5.95</b>
<b>M3</b>	<b>M1 + LogFreq</b>	<b>96,956</b>	<b>Yes</b>	Concreteness	- 20.71	4.74	- 4.37
M4	M3 + Length	[No entry]	No	ON	16.26	4.62	3.52
<b>M5</b>	<b>M3 + ON</b>	<b>96,949</b>	<b>Yes</b>	Arousal	- 13.3	4.57	- 2.91
M6	M5 + Valence	[No entry]	Yes				
<b>M7</b>	<b>M5 + Arousal</b>	<b>96,937</b>	<b>Yes</b>				
M8	M7 + Dominance	[No entry]	No				
<b>M9</b>	<b>M7 + Concreteness</b>	<b>96,923</b>	<b>Yes</b>				
<b>M10</b>	<b>M9 + CDV-Distance</b>	<b>96,886</b>	<b>Yes</b>				

Accepted model steps are shown in bold.

models show very strong effects of CDV-Diff in the hypothesized directions, as is shown in Fig. 14. Words with a large negative CDV-Diff (i.e., words closer to the noun than to the verb CDV) are recognized as nouns much more accurately than words with a large positive CDV-Diff (i.e., words closer to the verb than to the noun CDV). As expected, this effect is reversed for verbs. Words with a large CDV-Diff are recognized more accurately as verbs than are words with a small CDV-Diff. At the extremes, high-magnitude values of CDV-Diff are associated very consistently with incorrect decisions. Strongly noun-biased words, such as *mushrooms*, *diapers*, *homes*, and *bins*, were recognized as nouns 100% of the time, but as verbs less than 20% of the time. Similarly, strongly verb-biased words, such as *throbs*, *finds*, *builds*, and *snuggles*, were recognized with high accuracy as verbs, but rarely or never recognized as nouns.

We again analyzed the correct RT data using LME modeling with normalized predictors. CDV-Diff was the strongest predictor of RTs in both experiments.

The model analysis and fixed effects for noun decision are summarized in Table 10. The best model before entering CDV-Diff (AIC = 40,073) included random effects of stimulus and participant and fixed effects of logged word frequency, valence, and concreteness. Adding CDV-Diff improved the model substantially (AIC = 40,046). The relationship between the estimated RT and CDV-Diff is shown in Fig. 15.

The model analysis and fixed effects for verb decision are summarized in Table 11. The best model before entering CDV-Diff as a predictor (AIC = 48,066) included random effects of stimulus and participant and fixed effects of length, valence, arousal, and dominance. Adding CDV-Diff knocked out the last three predictors while improving the model very substantially (AIC = 47,984). The relationship between the estimated RT and CDV-Diff is shown in Fig. 15. Verbs are recognized more quickly as they are closer to the verb CDV and farther from the noun CDV.

## Discussion

The results from Experiments 3 and 4 provide a strong test of the hypothesis that human word category decisions are sensitive to distance from a CDV. Both experiments showed clear effects of CDV-Diff on accuracy and decision time, in the directions hypothesized. The effects are large. For noun decisions they are 50 ms per *SD*, or about 175 ms over the range of the stimuli. For verb decisions they are about 90 ms per *SD*, or about 270 ms over the range of the stimuli. The fact that we used exactly the same strings for both the noun and verb decisions in these experiments serves to verify that these large effects cannot be an artifact of some uncontrolled or poorly controlled lexical variable.

## General discussion

How are we to interpret these effects? What does it mean to be a “more typical” noun or verb? It is not immediately obvious how to answer these questions just from examining the closest neighbors. For example, the closest words to the plural-noun CDV are *bicycles*, *birds*, *campsites*, *gardeners*, *canoes*, *critters*, *scarecrows*, *birders*, *bobcats*, *volunteers*, and *garages*, a list that may seem disparate. A reasonable hypothesis is that the best exemplars of a class are those that are most similar to other members of that class, just as a robin is a better exemplar of a bird than is an ostrich, because a robin is more similar to other birds than is an ostrich. To test this hypothesis quantitatively, we correlated each word’s distance from its CDV with the number of words in common between its own closest 1,000 neighbors and the closest 1,000 neighbors of its CDV. The number 1,000 was chosen arbitrarily, but we have confirmed that other numbers of similar magnitude show similar effects. As is shown in Fig. 16, across the 500 words used in Experiments 1 and 2, the number of common neighbors (how much a word is like other words in its class) is a very strong negative predictor of distance from its CDV (for verbs:  $r = -.94$ ,  $p = 1.0E-234$ ; for nouns:  $r = -.84$ ,  $p = 2.5E-134$ ). The word “birds” is the nouniest noun<sup>3</sup> because there are many (300) words in common between that word’s closest 1,000 neighbors and the closest 1,000 neighbors of the plural-noun CDV.

## Conclusion

Skip-gram matrices constitute a very dense source of linguistic information. Most effort in psychology has concentrated on demonstrating that the matrices contain semantic information—that is, information about the meanings of words. As we noted above, several researchers have pointed out that there is also a great deal of information about the morphological and syntactic *class* of the word, by looking at vector offsets in word pair analogies (i.e., *fit:fitted::hope:X*). We have here demonstrated that a small set of simple operations (comparison to one or two vectors that can be deduced directly from the matrix) with just a single free parameter (number of words in the first vector by which to sort by the second vector) can account for a large portion of the variance in three important linguistic operations: recognizing parts of speech, recognizing morphological families, and performing different types of affixation. Our method is essentially an idealization of the vector offset method. We did not use vector subtraction to get the relevant offset but rather averaged across vectors that

<sup>3</sup> However, we noticed after all experiments were finished that it is actually classified erroneously, since *birds* is also a verb, though very rarely used in that form.

**Table 8** Model analysis (left) and final fixed effects (right) for LME model predicting accuracy in the ambiguous noun judgment task (Exp. 3), with normalized predictors

Model	Specification	AIC	Accept?	Fixed effects	Estimate	SE	z	p
<b>M0</b>	<b>(1   STIM)</b>	<b>4,064</b>	<b>Yes</b>	(Intercept)	1.34	0.21	6.51	7.43E-11
<b>M1</b>	<b>M0 + (1   ID)</b>	<b>3,742</b>	<b>Yes</b>	<b>CDV-Diff</b>	<b>-1.33</b>	<b>0.11</b>	<b>-12.54</b>	<b>&lt;2E-16</b>
M2	M1 + (1   ORDER)	[No convergence]	No	Concreteness	0.34	0.1	3.32	0.00091
<b>M3</b>	<b>M1 + LogFreq</b>	<b>3,723</b>	<b>Yes</b>	LogFreq	0.34	0.08	4.23	2.31E-05
M4	M3 + Length	[No entry]	No					
<b>M5</b>	<b>M3 + ON</b>	<b>3,718</b>	<b>Yes</b>					
<b>M6</b>	<b>M5 + Valence</b>	<b>3,706</b>	<b>Yes</b>					
M7	M6 + Arousal	[No convergence]	No					
<b>M8</b>	<b>M6 + Dominance</b>	<b>3,667</b>	<b>Yes</b>					
M9	M8 + Concreteness	[Knocked out ON]	No					
<b>M10</b>	<b>M3 + Valence + Dominance + Concreteness</b>	<b>3,580</b>	<b>Yes</b>					
M11	M10 + CDV-Diff	[Knocked out dominance and valence]	No					
<b>M12</b>	<b>M3 + Concreteness CDV-Diff</b>	<b>3,460</b>	<b>Yes</b>					

Accepted model steps are shown in bold.

characterize the word type of interest to get a characteristic vector (CDV for that word type). The vector offset method that has been popular in computer science is of less interest to psychologists, since it depends upon making analogies only between individual word pairs. Although vector averaging itself may debatably be psychologically implausible, the function that it emulates (category recognition by similarity) is clearly plausible, and, indeed, is assumed already by researchers who have used co-occurrence models to model semantics. The fact that we introduced no new operations in making the step from semantic categorization to part-of-speech categorization is scientifically satisfying since it reduces apparently disparate aspects of language to the same single operation.

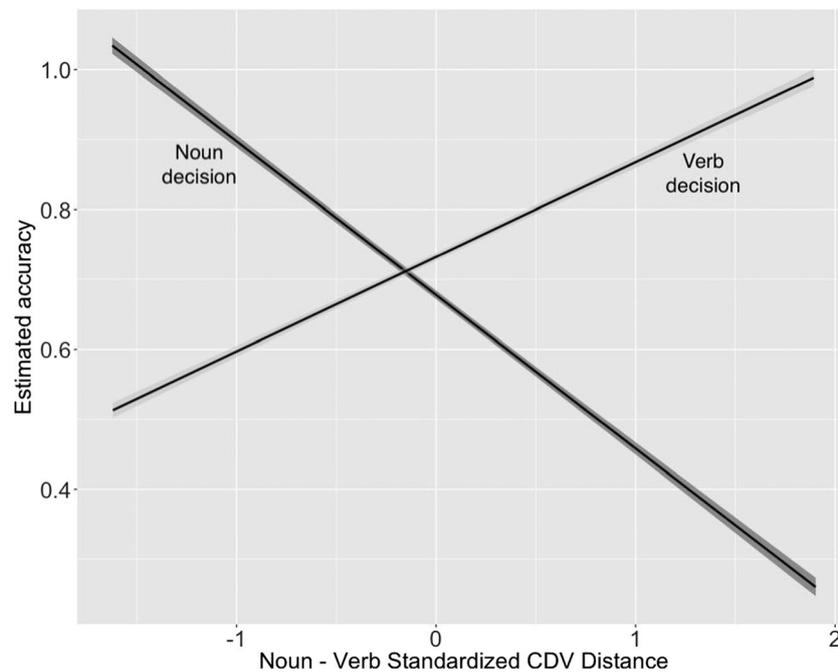
One problem that complicates decompositional approaches to dealing with morphology is the problem of pseudo-affixes, in which words either have a semantically transparent affix with no root (*grocer*) or seem to have a semantically transparent affix that is in fact not an affix (e.g., *corner*). This problem disappears in a purely semantic account of morphology. Our CDV for ACTOR+er puts *greengrocer* and *grocer* well inside the first decile of the category's words (1745th and 2673th, respectively, among all 78,278 words in a category with 3,841 exemplars). It puts *corner* much farther from the CDV, at the 13,168th position.

Others have pointed out that there are problems with the classical view of decompositional morphology. In their article showing that they could not replicate masked priming effects

**Table 9** Model analysis (left) and final fixed effects (right) for binomial LME model predicting accuracy in the ambiguous verb judgment task (Exp. 4), with normalized predictors

Model	Specification	AIC	Accept?	Fixed effects	Estimate	SE	z	p
<b>M0</b>	<b>(1   STIM)</b>	<b>4,626</b>	<b>Yes</b>	(Intercept)	1.51	0.18	8.39	<2e-16
<b>M1</b>	<b>M0 + (1   ID)</b>	<b>4,311</b>	<b>Yes</b>	<b>CDV-Diff</b>	<b>0.85</b>	<b>0.09</b>	<b>9.29</b>	<b>&lt;2e-16</b>
M2	M1 + (1   ORDER)	4,313	No	Arousal	0.23	0.1	2.36	0.018
M3	M1 + LogFreq	[No entry]	No	Dominance	0.2	0.09	2.28	0.023
M4	M1 + Length	4,308	No					
M5	M1 + ON	[No entry]	No					
M6	M1 + Valence	[No entry]	No					
<b>M7</b>	<b>M1 + Arousal</b>	<b>4,275</b>	<b>Yes</b>					
<b>M8</b>	<b>M7 + Dominance</b>	<b>4,259</b>	<b>Yes</b>					
<b>M9</b>	<b>M8 + Concreteness</b>	<b>4,249</b>	<b>Yes</b>					
M10	M9 + CDV-Diff	[Knocked out concreteness]	No					
<b>M11</b>	<b>M8 + CDV-Diff</b>	<b>4,184</b>	<b>Yes</b>					

Accepted model steps are shown in bold.



**Fig. 14** Estimated accuracies at noun and verb judgment (Y-axis) in Experiments 3 and 4, as a function of the standardized difference in distance from each category's CDV (X-axis), with 95% confidence intervals

previously attributed to obligatory morphological decomposition (after analyzing the prime as a random effect, rather than a fixed effect as usual), Milin, Feldman, Ramscar, Hendrix, and Baayen (2017) pointed out four problems with decompositional morphology (p. 37). One is that the token frequencies of pseudo-affixed words (such as *corner*) are often larger than the token frequencies of morphologically complex words, which would make mandatory decomposition an inefficient strategy. Similarly, the fact that most affixed words have diverse semantic connotations that are not predictable from their constituents would also suggest

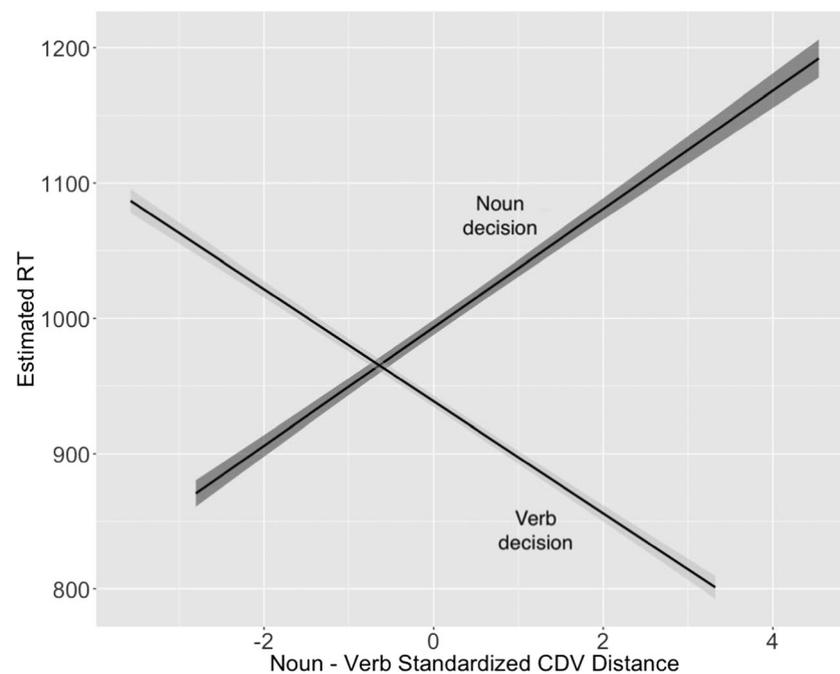
that decomposition was inefficient. Milin et al. discussed the morpheme *-er* as an example:

The problem is that *er* does not have a single semantic function, but instead expresses a wide range of meanings, including comparison (*greater*), agency (*walker*), instrument (*opener*), causation (*howler*), and patiency (*shooter*, as in *this bear is a shooter*; see Bauer, Lieber, & Plag, 2013; Booij, 1986). Furthermore, *er* is also found in words that are clearly subject (agent) nouns, but are not derived from any particular base words; for example,

**Table 10** Model analysis (left) and final fixed effects (right) for LME model predicting RT in the ambiguous noun judgment task (Exp. 3), with normalized predictors

Model	Specification	AIC	Accept?	Fixed effects	Estimate	SE	<i>t</i>
<b>M0</b>	<b>(1   STIM)</b>	40,556	Yes	(Intercept)	1,014.57	25.91	39.16
<b>M1</b>	<b>M0 + (1   ID)</b>	40,207	Yes	<b>CDV-Diff</b>	<b>50.29</b>	<b>10.33</b>	<b>4.87</b>
M2	M1 + (1   ORDER)	40,208	No	Concreteness	-46.67	10.23	-4.56
<b>M3</b>	<b>M1 + LogFreq</b>	<b>40,169</b>	<b>Yes</b>	LogFreq	-43.88	8.22	-5.33
M4	M3 + Length	[No entry]	No	Valence	-16.92	8.25	-2.05
M5	M3 + ON	[No entry]	No				
<b>M6</b>	<b>M3 + Valence</b>	<b>40,151</b>	<b>Yes</b>				
<b>M7</b>	<b>M6 + Arousal</b>	<b>40,129</b>	<b>Yes</b>				
M8	M7 + Dominance	[No entry]	No				
M9	M7 + Concreteness	[Knocked out arousal]	No				
<b>M10</b>	<b>M6 + Concreteness</b>	<b>40,073</b>	<b>Yes</b>				
<b>M11</b>	<b>M10 + CDV-Diff</b>	<b>40,046</b>	<b>Yes</b>				

Accepted model steps are shown in bold.



**Fig. 15** Estimated RTs for noun and verb judgment (Y-axis) in Experiments 3 and 4, as a function of the standardized difference in distance from each category's CDV (X-axis), with 95% confidence intervals

*father, mother and brother*, which fit in with the category of persons denoted by *er* in agent nouns such as *baker* and *butcher*, or *buyer* or *seller* (but do not fit the category of comparatives). This semantic fragmentation of *er* and the token-wise preponderance of pseudosuffixed words with *er* such as *her*, diminish the utility of a purely orthographic form representation for *er*. (p. 5)

Third, Milin et al. (2017) pointed out that evidence suggests that even nonmorphological embedded constituents in a word (such as *hat* in the word *hatch* or *ram* in the word

*drama*) have their semantics activated during lexical access (e.g., Baayen, Wurm, & Aycocck, 2007; Bowers, Davis, & Hanley, 2005).

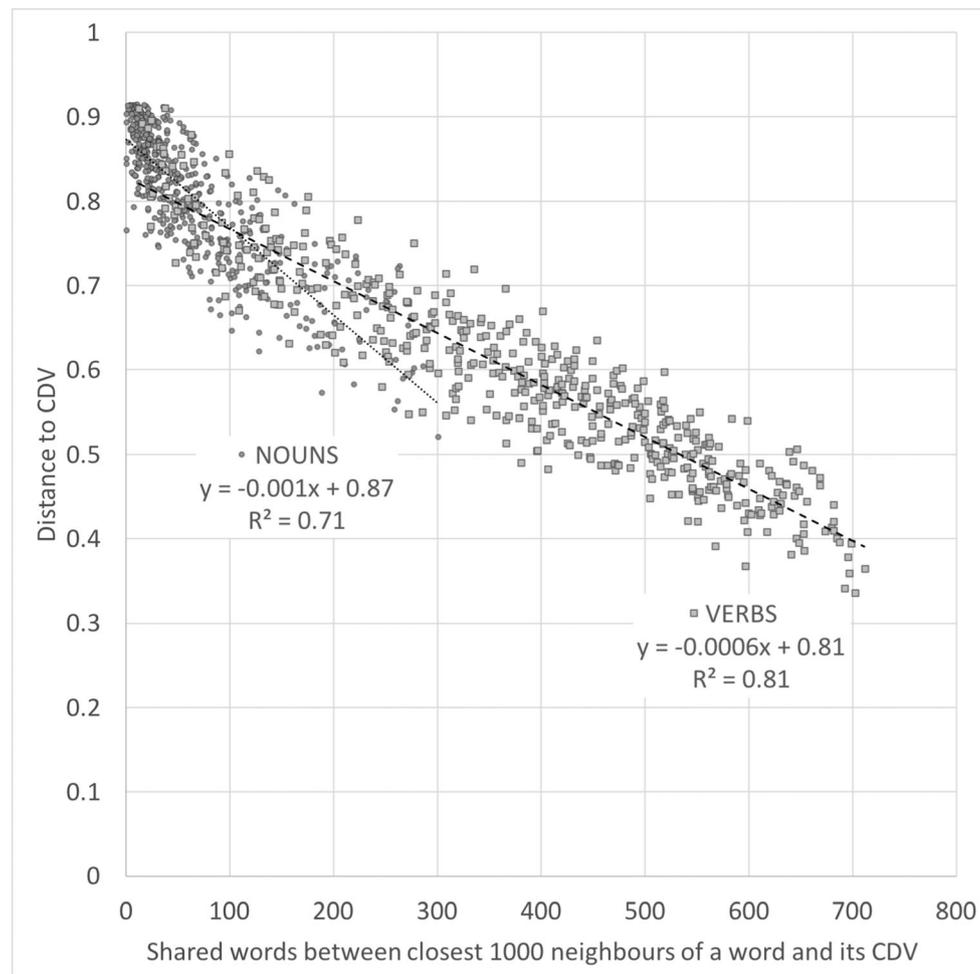
Our method also has no problem with homonyms, which are often eliminated when using the vector offset method because they introduce problematic noise. Since our measure can quantify distance to different CDVs, it can classify words that belong to more than one word class as being better or worse exemplars of each class.

It may seem that there is no simple semantic explanation for our well-documented ability to understand the meaning of

**Table 11** Model analysis (left) and final fixed effects (right) for LME model predicting RT in the ambiguous verb judgment task (Exp. 4), with normalized predictors

Model	Specification	AIC	Accept?	Fixed effects	Estimate	SE	t
<b>M0</b>	<b>(1   STIM)</b>	<b>48,500</b>	<b>Yes</b>	(Intercept)	950.96	22.79	41.73
<b>M1</b>	<b>M0 + (1   ID)</b>	<b>48,156</b>	<b>Yes</b>	<b>CDV-Diff</b>	<b>-90.5</b>	<b>6.64</b>	<b>-13.62</b>
M2	M1 + (1   ORDER)	48,158	No	Length	36.42	6.6	5.52
M3	M1 + LogFreq	[No entry]	No				
<b>M4</b>	<b>M1 + Length</b>	<b>48,135</b>	<b>Yes</b>				
M5	M3 + ON	[No entry]	No				
<b>M6</b>	<b>M4 + Valence</b>	<b>48,130</b>	<b>Yes</b>				
<b>M7</b>	<b>M6 + Arousal</b>	<b>48,101</b>	<b>Yes</b>				
<b>M8</b>	<b>M7 + Dominance</b>	<b>48,066</b>	<b>Yes</b>				
M9	M3 + Concreteness	[No entry]	No				
M10	M6 + CDV-Diff	[Knocked out dominance, arousal, and valence]	No				
<b>M11</b>	<b>M4 + CDV-Diff</b>	<b>47,984</b>	<b>Yes</b>				

Accepted model steps are shown in bold.



**Fig. 16** Relationship between distance to the plural-noun and verb CDVs and the number of words in common between the 1,000 closest neighbors of each word and of its CDV

novel affixed forms, such as the lexically unnecessary form *bluous*—that is, the root *blue* affixed with the adjective-forming suffix *-ous*. Since, as we discussed in the introduction, co-occurrence models define meaning entirely by a word’s proximity to related words, it suffices for fixing the meaning of a novel form to be able to select its correct related words. This can be done using the same methods we have used above to affix words: Simply take the closest neighbors to a target word, and then take the closest words among those neighbors to the CDV. If we do this for *bluous*, sorting the closest 100 neighbors of the word *blue* by their proximity to the CDV for the suffix *-ous*, the closest 20 words include not only the *bluous* synonym *bluish* (in the 19th position), but also *orangey*, *pinkish*, *blackish*, *purplish*, *reddish*, *greyish*, *greenish*, *orangish*, and *brownish*. Note that none of these contain the suffix we sorted by, but all contain a closely related adjective-forming suffix, allowing the inference from simple vector operations that *bluous* must be like *greenish*, but having to do with the color blue. Although more work will be required to see whether this method can model human intuitions or behaviors and whether there are parameters that make

it tractable in all cases, we believe it demonstrates that there is in principle a method to go from words to affixed pseudowords using a purely semantic analysis, and moreover, one that uses exactly the same operations that can demonstrably be used to both find affixed forms of root words and identify semantic categorical coherence.

Although we achieved very good success in classifying words (90% to 100% in selecting 1,000 exemplars of some word types [Table 1], and 49% to 100% success in categorizing 500 random exemplars from each of five word classes), we believe we have only set a lower bound on the performance of co-occurrence models in classifying word by morphological family and part of speech. We deliberately made no attempt to systematically optimize our CDVs, keeping them to 100 random exemplars in order to sidestep any criticisms of overfitting our models. It seems very likely that a more inclusive definition of those vectors would improve our performance on many of the classification tasks we have considered.

Limiting ourselves to at most two vectors in our automatic affixation examples is also a strong and unrealistic limitation. If we had allowed other evidence to weigh in on our affixation

operations (for example, negatively weighting words belonging to the class of words ending in *ing* or to the infinitive verb class as potential exemplars of the past tense) it is likely that we could improve our true positive rate substantially. We did not do so here because we did not wish to hand-fit our models in an *ad hoc* fashion, except insofar as we optimized the one free parameter,  $N$ . However, it seems likely that in a highly parallel system like the human brain, synthesizing multiple relevant dimensions in a more inclusive way than we have modeled here using just two dimensions is the norm.

Critics may point out that the Google news matrix we used to define our matrix is very large, containing as it does many more words than a person would be exposed to in their lifetime. We have offered our work here mainly as a proof of concept, to show that it is possible to extract part of speech and morphological information from co-occurrence alone. It is a separate issue to disentangle exactly how much co-occurrence information is enough to enable this extraction. We have not attempted to do so here, although we suspect that a much smaller corpus than the Google corpus would suffice. We also note that children have access to a great deal more relevant discriminatory information than merely what is contained in words. For example, they can get disconfirming evidence about the meaning of words from real-world experience, learning that a cat is not a dog or that an avocado is not a pear by direct experience with cats, dogs, avocados, and pears.

We have shown here that it is possible to treat part-of-speech identification the same way that semantic classification is treated in predict co-occurrence models, whose computational mechanisms are closely related to well-studied and simple animal learning models. We have also presented evidence from four experiments showing that humans are sensitive across the range of the continuous word-class cues derived from such models. We suggest that it is possible that the problem of word class identification is delimited in part by the same mechanisms that delimit semantic category identification, an interesting convergence of apparently disparate linguistic functions and simple learning models.

**Author note** This work was made possible by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

## References

- Abka, A. F. (2016). Evaluating the use of word embeddings for part-of-speech tagging in Bahasa Indonesia. In *2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA)* (pp. 209–214). Piscataway, NJ: IEEE Press.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Amon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*, 292–305.
- Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension on the basis of naive discriminative learning. *Psychological Review, 118*, 438–481. <https://doi.org/10.1037/a0023851>
- Baayen, R. H., Wurm, L. H., & Aycocock, J. (2007). Lexical dynamics for low-frequency complex words: A regression study across tasks and modalities. *Mental Lexicon, 2*, 419–463.
- Baker, C. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry, 10*, 533–581.
- Baroni, M., Dinu, G., & Kruszewski, G. (2014). *Don't count, predict!* A systematic comparison of context-counting versus context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 238–247). Stroudsburg, PA: Association for Computational Linguistics.
- Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1183–1193). Stroudsburg, PA: Association for Computational Linguistics.
- Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford reference guide to English morphology*. Oxford, UK: Oxford University Press.
- Booij, G. E. (1986). Form and meaning in morphology: The case of Dutch “agent nouns”. *Linguistics, 24*, 503–517.
- Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Automatic semantic activation of embedded words: Is there a “hat” in “that”? *Journal of Memory and Language, 52*, 131–143. <https://doi.org/10.1016/j.jml.2004.09.003>
- Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. Hayes (Ed.), *Cognition and the development of language* (pp. 11–54). New York, NY: Wiley.
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904–911. <https://doi.org/10.3758/s13428-013-0403-5>
- Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 117–156). Mahwah, NJ: Erlbaum.
- Chomsky, N. (1980). *Rules and representations*. Oxford, UK: Blackwell.
- Cotterell, R., & Schütze, H. (2015). Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1287–1292). Stroudsburg, PA: Association for Computational Linguistics.
- Drozd, A., Gladkova, A., & Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: Beyond king – man + woman = queen. In *Proceedings of COLING 2016: Technical articles* (pp. 3519–3530). Stroudsburg, PA: Association for Computational Linguistics.
- Durda, K., & Buchanan, L. (2008). WINDSOR: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods, 40*, 705–712. <https://doi.org/10.3758/BRM.40.3.705>
- Gladkova, A., Drozd, A., & Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL Student Research Workshop* (pp. 8–15). Stroudsburg, PA: Association for Computational Linguistics.
- Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review, 23*, 1744–1756.
- Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating human judgments from skip-gram vector representations of word meaning.

- Quarterly Journal of Experimental Psychology*, 70, 1603–1619. <https://doi.org/10.1080/17470218.2016.1195417>
- Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1–37. <https://doi.org/10.1037/0033-295X.114.1.1>
- 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. <https://doi.org/10.1037/0033-295X.104.2.211>
- Lazaridou, A., Marelli, M., Zamparelli, R., & Baroni, M. (2013). Compositionally derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: Vol. 1. Long Articles* (pp. 1517–1526). Stroudsburg, PA: Association for Computational Linguistics.
- Lin, C. C., Ammar, W., Dyer, C., & Levin, L. (2015). *Unsupervised POS induction with word embeddings*. arXiv preprint. arXiv:1503.06760
- Ling, W., Dyer, C., Black, A., & Trancoso, I. (2015). Two/too simple adaptations of Word2Vec for syntax problems. In *Human language technologies: The 2015 Annual Conference of the North American Chapter of the ACL* (pp. 1299–1304). Stroudsburg, PA: Association for Computational Linguistics.
- Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28, 203–208. <https://doi.org/10.3758/BF03204766>
- Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity on the basis of prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, 57–78. <https://doi.org/10.1016/j.jml.2016.04.001>
- Marcus, G. (1993). Negative evidence in language acquisition. *Cognition*, 46, 53–85.
- Marcus, G., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). Over-regularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57, 1–165.
- Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review*, 122, 485–515. <https://doi.org/10.1037/a0039267>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint. arXiv:1301.3781
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119). Cambridge, MA: MIT Press.
- Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751). Stroudsburg, PA: Association for Computational Linguistics.
- Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PLoS ONE*, 12, e0171935. <https://doi.org/10.1371/journal.pone.0171935>
- Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla–Wagner model. *Psychological Bulletin*, 117, 363–386. <https://doi.org/10.1037/0033-2909.117.3.363>
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1978). *The measurement of meaning*. Urbana, IL: University of Illinois Press. (Original work published 1957)
- Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.
- Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.
- Pinker, S. (2004). Clarifying the logical problem of language acquisition. *Journal of Child Language*, 31, 949–953.
- Ramscar, M., Dye, M., Gustafson, J. W., & Klein, J. (2013). Dual routes to cognitive flexibility: Learning and response-conflict resolution in the Dimensional Change Card Sort task. *Child Development*, 84, 1308–1323.
- Ramscar, M., Dye, M., & Klein, J. (2013). Children value informativity over logic in word learning. *Psychological Science*, 24, 1017–1023. <https://doi.org/10.1177/0956797612460691>
- Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language*, 89, 760–793.
- Ramscar, M., Dye, M., Popick, H. M., & O'Donnell-McCarthy, F. (2011). The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PLoS ONE*, 6, e22501. <https://doi.org/10.1371/journal.pone.0022501>
- Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, 34, 909–957.
- Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Vol. 1, pp. 133–142). Stroudsburg, PA: Association for Computational Linguistics.
- Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43, 151–160. <https://doi.org/10.1037/0003-066X.43.3.151>
- Rescorla, R. A. (2008). Rescorla–Wagner model. *Scholarpedia*, 3:2237, revision #91711. Accessed May 11, 2018, at [http://www.scholarpedia.org/article/Rescorla-Wagner\\_learning\\_rule](http://www.scholarpedia.org/article/Rescorla-Wagner_learning_rule)
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.
- Roberts, M. A., & Chater, N. (2008). Using statistical smoothing to estimate the psycholinguistic acceptability of novel phrases. *Behavior Research Methods*, 40, 84–93. <https://doi.org/10.3758/BRM.40.1.84>
- Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton* (Project Para Report No. 85-460-1). Ithaca, NY: Cornell Aeronautical Laboratory (CAL).
- Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, 44, 493–515.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928. <https://doi.org/10.1126/science.274.5294.1926>
- Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70, 27–52.
- Shaoul, C., & Westbury C. (2006). USENET orthographic frequencies for 1,618,598 types (2005–2006). Edmonton, AB: University of Alberta. Retrieved from <http://www.psych.ualberta.ca/~westburylab/downloads/>
- Shaoul, C., & Westbury, C. (2008). Performance of HAL-like word space models on semantic clustering. In M. Baroni, S. Evert, & A. Lenci (Eds.), *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics: Bridging the gap between semantic theory and computational simulations* (pp. 42–46).
- Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, 42, 393–413. <https://doi.org/10.3758/BRM.42.2.393>

- Shaoul, C., & Westbury, C. (2011). HiDEx: The high dimensional explorer. In P. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 230–246). Hershey, PA: IGI Global.
- Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla–Wagner model. *Psychonomic Bulletin & Review*, 3, 314–321.
- Sutton, R., & Barto, A. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.
- Ward, G. (1996). Moby project: Part of Speech List. Retrieved on September 7, 2018, from <https://archive.org/details/mobypartofspeech03203gut>
- Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. In *WESCON Convention Record Part IV* (pp. 96–104). New York, NY: Institute of Radio Engineers.