Psychon Bull Rev (2016) 23:1744–1756
DOI 10.3758/s13423-016-1053-2

THEORETICAL REVIEW

The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics

Geoff Hollis[1] · **Chris Westbury**[1]

Published online: 2 May 2016

Abstract Notable progress has been made recently on computational models of semantics using vector representations for word meaning (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). As representations of meaning, recent models presumably hone in on plausible organizational principles for meaning. We performed an analysis on the organization of the skip-gram model's semantic space. Consistent with human performance (Osgood, Suci, & Tannenbaum, 1957), the skip-gram model primarily relies on affective distinctions to organize meaning. We showed that the skip-gram model accounts for unique variance in behavioral measures of lexical access above and beyond that accounted for by affective and lexical measures. We also raised the possibility that word frequency predicts behavioral measures of lexical access due to the fact that word use is organized by semantics. Deconstruction of the semantic representations in semantic models has the potential to reveal organizing principles of human semantics.

Keywords Semantics · Affect · Skip-gram · Language · Semantic differential · Principal components analysis · Meaning · Emotion · Co-occurrence

One of the few surviving documents attributed to the Pythagoreans (in this case by Aristotle, in his *Metaphysics*, trans. W. D. Ross, 1924) is a list of opposing principles. This list consists of ten pairs of opposites posited as basic organizing principles, such as "finite/infinite," "right/left," and "rest/motion." It is probably fair to say that this is the oldest formal psychological theory in the Western intellectual tradition. Since Aristotle's times, many efforts have been made to identify the basic dimensions of meaning. In this article, we take a new look at this problem using statistical methods.

One of the first modern dimensional approaches to semantics was the semantic differential, pioneered by Osgood, Suci, and Tannenbaum (1957). They asked participants to rate stimuli along a variety of bipolar axes, such as "dirty/clean," "good/bad," or "big/small." The ratings were factor-analyzed by Osgood et al. to identify systematic relations. They reported that a small number of factors regularly accounted for a large portion of the variance in the ratings, and that these factors had interpretable factor loadings. The primary factors regularly had to do with the pleasantness of the concept (*evaluation*), its energetic potential (*activity*), and the degree to which it could affect change (*potency*). More recently, the terms of *valence*, *arousal*, and *dominance* have been adopted. This change in terminology is in alignment with the notion that these factors are affective in nature (Mehrabian, 1996; Warriner, Kuperman, & Brysbaert, 2013).

An interesting aspect of Osgood et al.'s (1957) work was that these three factors consistently arose, regardless of stimulus type. They have been seen for judgments of words, paintings, sculptures, and sonar signals (Osgood et al., 1957), and more comprehensively in the connotative aspects of color (e.g., Fang, Murumatsu, & Matsui, 2015; Ou, Luo, Woodcock, & Wright, 2004a, 2004b). This result suggests that it is productive to discuss semantics as a continuous, dimensioned space, with *semantic distinctions* that are largely aligned with *affective dimensions*.

✉ Geoff Hollis
hollis@ualberta.ca

[1] Department of Psychology, University of Alberta, T6G 2E9 Edmonton, Alberta, Canada

Although semantic differential studies suggest that affect is the most robust determinant of meaning, Osgood et al. (1957) were careful to note that semantics is not *only* based on affect, as other types of factors (often less easily interpretable) do emerge when applying the semantic differential. Furthermore, the relative importances of valence, arousal, and dominance vary with the evaluative context. Complex evaluative contexts may even reshape semantic space, causing the main dimensions to become dependent on (i.e. correlated with) one another. For example, Osgood et al. demonstrated that evaluation, activity, and potency are not orthogonal dimensions within political evaluative contexts. Political concepts that are *good* also tend to be *active* and *potent*. In line with this, the semantic differential occasionally reveals hybrid dimensions (e.g., an *activity* plus *potency* dimension, dubbed *dynamism*).

Applications of the semantic differential as a tool for studying semantics have waned in recent years. However, the idea that semantics can be expressed as a dimensioned space has not. Recent work with co-occurrence models of lexical semantics can be viewed as a technologically updated extension of much of Osgood's earlier work (e.g., Durda & Buchanan, 2008; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996; Mikolov, Chen, Corrado, & Dean, 2013; Rohde, Gonnerman, & Plaut, 2006; Shaoul & Westbury, 2010). Although their technical details vary, all of these models work from the basic assumption that the surrounding context of a word is informative of its meaning. Most of them apply this basic assumption literally, representing word meaning as a vector of occurrences within contexts defined over documents of text (e.g., Landauer & Dumais, 1997), or as neighboring words within a sentence (e.g., Lund & Burgess, 1996). This approach lends itself to some interesting applications. For instance, the similarity of meaning between two words can be assessed by measuring the similarity between their co-occurrence vectors. Co-occurrence vectors contain enough information to pass tests for basic verbal ability (e.g., Landauer & Dumais, 1997), to accurately predict human judgments of valence and arousal (e.g., Hollis & Westbury, 2016; Mandera, Keuleers, & Brysbaert, 2015; Recchia & Louwerse, 2015; Westbury, Keith, Briesemeister, Hofmann, & Jacobs, 2015), and to account for behavioral effects of high-level lexical properties such as subjective familiarity (Westbury, 2014) and imageability (Westbury et al., 2013).

A recent co-occurrence model is Google's continuous skip-gram model (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Rather than representing word meaning as a co-occurrence vector, the skip-gram model trains a neural network to predict the surrounding contextual details of a sentence, given a prompt word. Each word's vector representation is an input to a neural network that is shaped through error back-propagation. Google's model is a major breakthrough in vector representations of semantics, performing substantially better on tests of semantic and syntactic understanding than alternate models (see Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013).

One characteristic that makes the skip-gram model unique is the density of its vectors. The vector representations in co-occurrence models are usually long and sparse, since they express relationships between one word and other words (of which there are many), and most of these cells are null (because most words do not co-occur). The skip-gram model uses shorter vector representations, in lengths of tens or hundreds of information-dense cells encoding predictions about anticipated context. This suggests that skip-gram vectors may represent semantics more transparently, picking up on semantic distinctions that would be directly relevant for predicting relationships between words and their contexts. It has been noted that skip-gram vectors produce interesting behavior under addition and subtraction. For instance, the vector for "King," minus the vector for "Man," plus the vector for "Woman" approximates the vector for "Queen." This is not an exceptional case but, rather, a consequence of how the skip-gram model and related models represent distinctions between semantically unrelated concepts (Levy & Goldberg, 2014a). The skip-gram model performs well at analogical reasoning across a broad range of topics (Mikolov, Yih, & Zweig, 2013). The skip-gram model represents meaning as vectors organized within a multidimensional space that preserves the linear relationships of human-intelligible semantic concepts.

Thus, the skip-gram model brings us full circle to Osgood's questions of whether meaning can be organized coherently in a dimensional space and, if so, what are the underlying dimensions of organization? The purpose of our work is to see whether coherent semantic dimensions can be reverse-engineered from the vector representations of meaning within the skip-gram model.

Method

We made use of precomputed skip-gram vectors released by Google (Word2vec, 2013), with vectors of length 300 trained using the skip-gram architecture on a small portion of the (proprietary) Google News corpus. For specific implementation details and parameter considerations, see Mikolov, Chen, Corrado, and Dean (2013) and Mikolov, Sutskever, Chen, Corrado, and Dean (2013).

An architecture similar to the skip-gram model is the continuous bag of words (CBOW) model (Mikolov, Chen, et al., 2013). Whereas the skip-gram architecture trains vector representations by predicting surrounding context from a single word, the CBOW architecture trains vector representations by predicting a word from the surrounding context. In comparisons of the skip-gram and CBOW architectures, both perform

similarly for capturing systematicities in syntactic relationships. However, the skip-gram architecture has superior performance for capturing semantic relationships (e.g., Qiu, Cao, Nie, Yu, & Rui, 2014; Mikolov, Chen, et al., 2013; Mnih & Kavukcuoglu, 2013). The skip-gram architecture is most appropriate for our purposes.

In anticipation of the skip-gram model organizing word meaning along semantic dimensions that psychology has identified as relevant to human cognition, we performed our analyses on the union of words that appear in recent crowdsourced norms of human judgments for valence, arousal, and dominance (Warriner et al., 2013), concreteness (Brysbaert, Warriner, & Kuperman, 2014), and age of acquisition (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). We also included lexical variables available from the English Lexicon Project (log frequency in the HAL corpus, orthographic neighborhood size, and word length; Balota et al., 2007). We used these databases because they have the largest coverage of the English language for the respective measurements contained within each. In total, we used 12,344 words in our analysis.

Results

Identification of semantic dimensions—Raw vectors

Our primary goal was to understand the semantic dimensions along which the skip-gram model organizes meaning. We anticipated finding indications that it organizes meaning along affective dimensions (e.g., *valence*, *arousal*, and *dominance*). We additionally considered the possibility that other semantic dimensions (e.g., *concreteness* or *age of acquisition*) or lexical variables that are known to impinge on lexical access (e.g., *log frequency*, *length*, or *neighborhood size*) may play a role in the organization of word meaning. If the skip-gram model organizes word meanings along dimensions corresponding to human-relevant semantic or lexical variables then, for each variable, there should be one primary vector dimension with which human judgments strongly correlate, relative to the rest of the vector dimensions, and few dimensions overall should be correlated with any particular variable.

We defined two formal criteria for labeling vector dimensions. A vector dimension was provided a semantic or lexical label, x, if (1) that dimension, d, was more strongly correlated with measurements of x than was any other dimension, and (2) measurements of no other semantic or lexical construct were more strongly correlated with d than were measurements of x.

These criteria were deliberately chosen not to require exclusive relationships between vector dimensions and the lexical and semantic variables being considered. Osgood et al. (1957) observed that in complex evaluative contexts, semantic dimensions have a tendency to rotate toward each other (i.e., to become nonorthogonal) and form dependencies. Skip-gram vectors were trained on a large set of unstructured, real-world texts pertaining to news-related topics. It seems likely that this would constitute the type of context where we might see evaluative dependencies forming between semantic dimensions. In such a situation, we might expect one dimension to be strongly related to multiple semantic or lexical measures (though multiple dimensions would not necessarily be related to the same semantic or lexical measure). Consequently, we employed criteria for providing interpretive labels that only required a variable and a dimension to be more strongly related to each other than they were to other variables or other dimensions.

We started by correlating each of our eight variables with each of the 300 dimensions in skip-gram word representations. Using a Bonferroni-corrected alpha of $a = .05/2{,}400$ (.00002), we found that 208 dimensions (69 %) were reliably correlated with *valence judgments*, 175 (58 %) with *arousal judgments*, 187 (62 %) with *dominance judgments*, 238 (79 %) with *concreteness judgments*, 181 (60 %) with *age-of-acquisition judgments*, 149 (50 %) with *log frequency*, 147 (50 %) with *orthographic neighborhood size*, and 155 (52 %) with *word length*. The largest-magnitude correlation observed was $r(12{,}342) = .39$, between concreteness judgments and vector dimension 80.

We observed many small but reliable correlations between semantic and lexical variables, on the one hand, and vector dimensions, on the other. These correlations were unlikely to be spurious, since we had used a conservative alpha correction. However, for comparison, we constructed a chance model by randomly shuffling the order of the values within each of our 12,344 word vectors and reran the analyses above. A single test, for concreteness, came out significant. The observed correlation was $r(12{,}342) = .04$.

We next looked for pairs of semantic/lexical variables and vector dimensions that were more strongly related to each other than to other variables/dimensions. For each variable, we identified the vector dimension to which it was most strongly related. We then verified that (1) that variable was statistically more strongly correlated with that dimension than with other dimensions (via a comparison of correlations using the Fisher r-to-z test with an alpha of $a = .05$) and (2) that dimension was more strongly correlated with that variable than any other variable (using the same test). By these criteria, only Dimension 80 received a semantic label (*concreteness*). Values along this dimension were correlated with concreteness judgments at $r = .39$. However, we note that 237 other dimensions (79 %) were also correlated with those judgments, of which five others had correlations at $r > .30$: Dimensions 41, 175, 295, 102, and 78 (rs = .35, .33, .32, .32, and .31, respectively).

We concluded that the space defined by the skip-gram model's raw vectors does not organize word meaning

according to the semantic or lexical variables we had tested. We based this conclusion on the facts that (1) each of our semantic and lexical variables was reliably correlated with most (50 %–79 %) of the vector dimensions and (2) although our criteria provided a label of *concreteness* to one dimension, other dimensions were reliably correlated with concreteness judgments, as well. We believe the more plausible interpretation of these results is that the skip-gram model simply uses information carried by these variables to make semantic distinctions without preserving an organization of semantic space onto which standard psycholinguistic variables will map.

Identification of semantic dimensions—Principal components

Levy and Goldberg (2014b) provided a proof that with sufficient training, the skip-gram model's word vector matrix is a factorization of the pointwise mutual information shared between words for which the model has learning representations and the contexts in which they are presented during training. We were unable to find any published work that addressed how many training epochs are required for such a convergence, possibly because the skip-gram algorithm arrives at high-quality vector representations in a single training epoch, as long as a large enough training corpus is used (Mikolov, Chen, et al., 2013). We were interested in testing whether further factorization of skip-gram vectors would change the relationship between vector dimensions and human judgments of semantics.

To test this, we performed principal component analysis (PCA) using singular-value decomposition on the 300 vector dimensions, after first mean-centering each dimension and scaling it to have unit variance. The R function *prcomp* was used to compute the PCA. It took 254 principal components (PCs) to account for 95 % of the variance between the vector dimensions, leaving 46 PCs to account for the remaining 5 %. The chance model required 280 PCs to account for 95 % of the variance between dimensions. We constructed 99 other chance models by randomizing values within each word vector. Each other chance model likewise required 280 PCs to account for 95 % of the variance between dimensions. The skip-gram model encodes word meaning in a highly nonredundant form, as we know from the fact that many PCs were required to account for the variance between dimensions. However, the 300 dimensions for representing word meaning were not entirely orthogonal. Our analysis of the chance models indicated that orthogonal dimensions would require approximately 280 PCs to account for 95 % of the variance between dimensions. Shuffling values within each word vector breaks dependencies between the vector dimensions, making them orthogonal.

We repeated our analysis of the raw vectors on the component scores extracted from the PCA. We again used a Bonferroni-corrected a of .05/2,400 to assess statistical significance. As compared to the raw vectors, fewer component scores were correlated with our variables: 38 (13 %) component scores were correlated with *valence judgments*, 37 (12 %) with *arousal judgments*, 35 (12 %) with *dominance judgments*, 23 with *concreteness judgments* (8 %), 73 (24 %) with *age of acquisition judgments*, 99 (33 %) with *log frequency*, 37 (12 %) with *orthographic neighborhood size*, and 40 (13 %) with *word length*.

We also constructed a chance model by performing PCA on the randomly shuffled vector values. Four comparisons for concreteness came out statistically reliable (PC74, r = –.044; PC116, r = –.038; PC227, r = .047; PC257, r = .037). No other comparisons were statistically reliable.

Information relevant to semantic and lexical variables is contained within fewer PCs than the raw vector dimensions. Following Osgood et al.'s (1957) finding that semantic distinctions tend to be made along affective dimensions, we hypothesized that valence judgments, arousal judgments, and dominance judgments would correlate highly with the earlier PCs that were extracted. Furthermore, we hypothesized that the earlier PCs would meet our criteria for being identified as an interpretable affective dimension, on the basis of our above labeling criterion. We had no hypotheses about the relationship between nonaffective variables and the extracted PCs.

In contrast to the raw vector representations, multiple PCs met our criterion for being labeled with interpretable semantic or lexical names. All of these PCs were extracted early by PCA. PC1 was labeled as *word frequency* (r = .42), PC2 as *concreteness* (r = .64), PC5 as *valence* (r = .50), and PC7 as *dominance* (r = .38).

PC4 correlated strongly with both age of acquisition (r = .36) and word length (r = .35). Because there was no reliable difference between these correlations (z = 0.51, p = .61), PC4 did not meet our criteria for receiving either label. We considered the possibility that PC4 might be organizing words along an axis of *specificity of meaning*: Longer words and words that are learned later in life are generally more specific in what they mean. With this in mind, we defined a measure of word specificity (*age of acquisition * word length*) and correlated it with the 300 PCs. It correlated most strongly with PC4 (r = .43), and this correlation was stronger than the one between PC4 and either age of acquisition judgments (z = 6.49, p = 8.58e–11) or word length (z = 7.38, p = 1.58e–13) alone. PC4 therefore met our criteria for being labeled a dimension of *meaning specificity*.

PC3 likewise had comparatively high correlations with age of acquisition judgments, word length, and orthographic neighborhood size, but did not meet our criteria for being labeled with an interpretable name. PC6, PC8, and all later PCs had nonsignificant or weak correlations with all variables and failed to meet our criteria for being given an interpretable semantic or lexical label. The correlation

magnitudes between the first eight PCs and our eight variables are presented in Table 1.

We draw two main conclusions from these results. The first is that the early PCs extracted from the skip-gram model display the tendency to organize word meaning according to a single lexical or semantic variable, whereas other PCs have a tendency not to be as strongly correlated with that same variable. Furthermore, two of these PCs (PC5 and PC7) organize word meaning along affective axes, supporting the point expressed by Osgood et al. (1957) that affect plays a central role in semantic distinctions.

Second, although some variables had a tendency to be more strongly correlated with one PC than with others, none of the variables were exclusively correlated with a single PC. Osgood et al. (1957) pointed out that in complex evaluative contexts, semantic dimensions have a tendency to rotate toward each other and form dependencies. It is possible we are witnessing the consequences of a complex evaluative context, given that the skip-gram model was trained on a large set of unstructured, real-world texts pertaining to news-related topics.

Face validity of the semantic dimensions

We visually inspected words with high and low loadings on PC1, PC2, PC4, PC5, and PC7 to check for the face validity of our interpretations of their organizing themes. The five words that loaded highest on PC2 (concreteness) were *cynical*, *pathetic*, *arrogant*, *laughable*, and *clueless* (Average [*SD*] concreteness judgment [out of 5]: 1.9 [0.18]). The other pole contained the words, *rotor*, *compressor*, *duct*, *infuser*, and *tubing* (Average [*SD*] concreteness judgment: 4.15 [0.54]). One pole contains affectively loaded social evaluations, whereas the other pole contains affectively neutral concrete objects. PC2 has face validity as a dimension of concreteness.

Table 1 Correlation strengths between eight lexical and semantic variables and the first eight principal components extracted from skip-gram vector representations

	Principal Component							
Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Log frequency	**.42**	.08	.07	.00	.03	.03	.12	.08
Concreteness	.39	**.64**	.03	.28	.11	.06	.08	.04
AoA	.12	.12	.17	**.36**	.00	.07	.08	.10
Length	.19	.07	.23	**.35**	.09	.03	.07	.04
Valence	.02	.21	.14	.08	**.50**	.05	.33	.00
Arousal	.03	**.27**	.10	.03	.05	.09	.15	.02
Dominance	.09	.14	.15	.12	.36	.02	**.38**	.00
Neighborhood size	.09	.05	.18	**.27**	.10	.06	.05	.01
AoA * Length	.21	.13	.24	**.43**	.06	.06	.08	.03

Bold numbers indicate the PCs that a particular variable is most strongly related to

The five words that loaded highest on PC4 (specificity of meaning) were *intracranial*, *temporal*, *physiological*, *pathological*, and *bodily* (Average [*SD*] AoA/length: 12.82 [2.98]/10.20 [3.03]). The five words that loaded lowest on PC4 were *clobber*, *grab*, *hang*, *hightail*, and *unload* (Average [*SD*] AoA/length: 7.98 [2.19]/5.80 [1.79]). The highest-loading words are technical and/or associated with academics, whereas the lowest-loading words are more colloquial in use. PC4 has face validity as a dimension of meaning specificity.

The five words that loaded highest on PC5 (valence) were *picturesque*, *splendid*, *magnificent*, *sparkling*, and *vibrant* (Average [*SD*] valence judgment [out of 7]: 6.5 [0.72]). The five words with the lowest loadings on PC5 were *harmful*, *rectal*, *anus*, *sphincter*, and *penis* (Average [*SD*] valence judgment: 3.4 [0.9]). The primary distinction between these words is one of valence.

The five words that loaded highest on PC7 (dominance) were *barren*, *hellish*, *desolate*, *horrific*, and *desolation* (Average [*SD*] dominance rating [out of 7]: 2.7 [0.8]), versus *communicator*, *articulate*, *conversationalist*, *compliment*, and *conversational* (Average [*SD*] dominance rating: 6.3 [0.7]). The first five capture a sense of "lack of control," whereas the latter set capture a sense of "being in control," which is how the construct of dominance is defined (e.g., Warriner et al., 2013). PC7 appears to have face validity as a dimension of dominance.

The five words that loaded highest on PC1 (word frequency) were *implement*, *evaluate*, *finalize*, *strengthen*, and *expedite* (Average [*SD*] log frequency: 7.72 [1.63]). The five words that loaded lowest on PC1 were *cherub*, *puss*, *wienie*, *hussy*, and *senorita* (Average [*SD*] log frequency: 4.82 [1.05]). It is not obvious from inspection of the poles that PC1 is organizing words by word frequency, despite the fact that the frequencies for the two sets are reliably different [Welch's $t(6.28) = -3.46$, $p = .01$]. Rather, the high-loading words all involve the sustained application of power, whereas the low-loading words have a tendency to be discussed as things that are "acted on" rather than "acting." These themes of activity and potency suggest that PC1 may be related to Osgood et al.'s (1957) hybrid affective concept of *dynamism*. However, it is also clear that one pole tends strongly toward femininity, suggesting that this pole may be organizing words along a feminine/masculine axis. These interpretations are not mutually exclusive: Feminist literature points out that general discourse conflates females as objects and males as subjects (Fredrickson & Roberts, 1997). Similar patterns were noted when observation was extended out to the top and bottom 20 words: the bottom pole contained additional words like *minx*, *wench*, *floozy*, *fatso*, *twat*, *pecker*, *sleepyhead*, *chick*, *sweetie*, and *dude*, whereas the top pole contained words like *consolidate*, *provide*, *prioritize*, *accelerate*, *establish*, *assess*, *improve*, *facilitate*, *defer*, and *extend*.

The idea that masculinity/femininity might be an organizing semantic axis in an old one. Whorf (1945) pointed out that

> smaller animals usually are "it"; larger animals often "he"; dogs, eagles, and turkeys usually "he"; cats and wrens usually "she," body parts and the whole botanical world "it"; countries and states as fictive persons (but not as localities) "she"; cities, societies and corporations as fictive persons "it"; the human body "it," a ghost "it"; nature "she"; watercraft with sail or power and named small craft "she"; unnamed rowboats, canoes, rafts, "it," etc. (p. 3)

Many empirical studies have supported the claim that gender plays a role in semantic processing. Osgood et al. (1957) showed that there was a strong relationship between their basic dimension of *potency* and judgments by English speakers of masculinity/femininity. Mismatches between expected (stereotyped) and marked gender (i.e., a butcher referred to in English with a feminine pronoun) have repeatedly been shown to impede comprehension in a variety of tasks (reviewed in Scheutz & Eberhard, 2004; more recently, see Bender, Beller, & Klauer, 2016; Esaulova & von Stockhausen, 2015).

The relationship between grammatical gender and biological (often called "natural" or "semantic") gender is complex and unclear (see the discussions in, e.g., Andonova, D'Amico, Devescovi, & Bates, 2004; Baron, 1971; Boroditsky, Schmidt, & Phillips, 2003; Konishi, 1993). This is in part because many languages mark gender differently and because grammatical and natural gender can interact. Konishi studied 54 masculine/feminine noun pairs that were oppositely gendered in Spanish and German, and found that the grammatical gender affected semantic judgments of *potency*. Boroditsky, Schmidt, and Phillips reported that their experimental participants were better at remembering proper names paired with nouns when the names and grammatical gender were gender-consistent. Other complications in understanding the relationship between grammatical and natural gender are that the strength of the correlation between grammatical and natural gender varies between languages (Andonova, D'Amico, Devescovi, & Bates, 2004), and the relationship often seems arbitrary (Mark Twain, 1880/1935, famously complained that "In German, a young lady has no sex, while a turnip does" [p. 259]). Nevertheless, the fact that many languages do mark gender grammatically suggests that the male/female distinction may anchor a basic semantic axis (see, e.g., Konishi, 1993; MacKay, 1999; Mills, 1986).

In contrast to gender, log frequency is not a semantic variable, so we believe it noteworthy that information correlated with log frequency is captured by a semantic model that never receives frequency information as input. We believe that PC1 being primarily correlated with log frequency is indicative of the fact that the need to convey specific types of meaning during communication organizes how words are used. Some topics will be discussed more frequently than others (e.g., societally relevant topics), meaning that variation across word frequency will necessarily be tied to variation across the semantic content of words. Examination of the words with high and low loadings on PC1 suggest that topics relating to Osgood et al.'s (1957) concept of *dynamism* and/or topics of *gender* are aspects of semantics organizing the frequency of word use. We discuss PC1 in more detail later on.

Unlabeled PCs

We now turn our attention to PC3 and PC6, both of which were earlier PCs that were unnamed by our criteria for labeling a semantic dimension.

The five words with the highest loadings on PC3 are *compress*, *glide*, *add*, *soak*, and *flatten*. The five words with the lowest loadings are *servant*, *businessman*, *policeman*, *lawyer*, and *journalist*. When observation was extended out to the top and bottom 20 words, the bottom words continued in the theme of personal titles, and the top words continued in the theme of actions. This dimension appears to be picking up the distinction between actions and actors and may be organizing words along a dimension of *agency*.

The five words with the highest loadings on PC6 are *mango*, *herbs*, *buckwheat*, *honey*, and *edible*. The bottom five words were *throttle*, *footwork*, *microphone*, *forearm*, and *teammate*. The pattern of edible and nonedible objects continued as the top and bottom 20 words were examined. PC6 may possibly be organizing words along a dimension of *edibleness*.

Of course, caution is necessary when interpreting the content of high- and low-loading words. By examining only a handful of words along a dimension that organizes over ten thousand, we may be basing interpretations on the pathological form of an underlying organizing principle (should one even exist). However, more detailed examination is not tractable, nor do we believe that it would provide strong empirical support for a meaningful interpretation of these dimensions. Consequently, we believe that such inspection can provide support for an a priori interpretation of a dimension (e.g., PC2, PC4, PC5, PC7), but is not strong grounds for making an interpretation of a dimension (e.g., PC1, PC3, PC6). Nonetheless, we believe such inspection may prove to be a useful basis for future experimental research on how humans organize semantic concepts.

These findings support the earlier conclusion that PC2 is organizing words along a dimension of *concreteness*, PC4 of *meaning specificity*, PC5 of *valence*, and PC7 of *dominance*. Additionally, these findings provide possible interpretations for PC1 (*dynamism and/or gender*), PC3 (*agency*), and PC6 (*edibleness*) that motivate further research on the extent to

which these concepts play a role in organizing the human mental lexicon.

Prediction of word association strength

If skip-gram PCs are psychologically plausible dimensions of semantics, we would expect them to have some predictive validity of word association strength. Nelson, McEvoy, and Schreiber (2004) provided a norms set for the association strengths between approximately 72,000 word pairs. Forward association strengths are based on the likelihood that one word (the target) will be elicited during free recall when prompted with the other word (the cue). A total of 56,344 cue–target pairs overlapped with the 12,344 words used in these analyses. We used the forward association strengths between these 56,344 cue–target pairs in our next analysis.

Overall, we found that word vector similarity (measured by cosine distance) between the non-PCA-transformed (raw) vectors reliably predicted forward association strength [$r(56, 342) = .27$, $p < 2.2e{-}16$]. A correlation of similar strength was observed [$r(56{,}342) = .27$, $p < 2.2e{-}16$] when the raw vectors were rotated with PCA before similarity measures were computed.

Comparisons of human associations and vector similarity based on cosine similarity are conservative, because cosine similarity treats each dimension as equally relevant when predicting human judgments. The possibility that some dimensions may matter more than others is not tested. We therefore repeated our attempts to predict forward association strength from the PCA-transformed word vectors using regression techniques instead.

Our first step was to create difference vectors for each word pair. Difference vectors were computed by taking the absolute differences between each dimension of the PCA-transformed cue-word vector and its corresponding dimension for the PCA-transformed target-word vector. We then individually regressed each dimension from the difference vectors on forward association strength. The results are presented in Fig. 1. Differences along PCs extracted earlier from the PCA predicted forward association strength better than differences along PCs extracted later. The drop-off of the contributions of PCs (from first to last) was fitted well by a linear function [$r(298) = -.86$, $p < 2.2e{-}16$]. When differences along all 300 PCs were included in a single regression equation, the model predictions correlated with forward association strength at $r = .30$ ($p < 2.2e{-}16$).

Some readers may be surprised by the small correlations observed when forward association strength was predicted from vector similarity or individual vector dimensions. This is not unique to the skip-gram model. Similarly small correlations have been observed when latent semantic analysis word similarity measures are used to predict forward association strength ($r = .27$; Nelson et al., 2004).

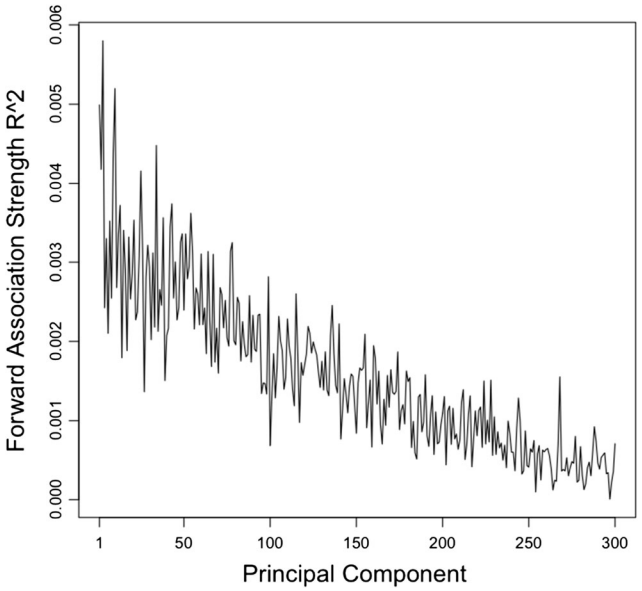

Fig. 1 Proportions of variance accounted for in forward association strength (Nelson, McEvoy, & Schreiber, 2004) by absolute differences between the cue and target along 300 principal components (PCs) extracted from skip-gram word vectors. The PCs extracted earlier account for more variance in forward association strength, suggesting that they have increased relevance to understand how humans make semantic distinctions

A few considerations related to how forward association strength is calculated make sense of these low correlations. Forward association strength is calculated on the basis of the likelihood that a particular word will be generated when cued by another word in a free-recall task. Such decisions are not driven exclusively by semantics. For instance, word rhyme influences calculations of associative strength (Maki, McKinley, & Thompson, 2004), as is the case when *car* elicits *bar*. Associative links also drive free recall (Nelson et al., 2004), as is the case when *basic* elicits *instinct* (due to learned associations from the movie *Basic Instinct*). Co-occurrence models like the skip-gram model provide an absolute measure of word similarity, whereas cued free recall provides a measure of relative word similarity. Measures of relative similarity create distortions in any sort of similarity landscape. Consider the situation in which *robin* or *ostrich* is used as a cue. Although *ostrich* is less similar to *bird* than *robin* is, *ostrich* may be no more similar to anything else than to *bird*, resulting in high forward associative strength anyway. Finally, since calculations of forward associative strength are based on cued free recall, norm sets like Nelson et al. (2004) are necessarily limited to only words that are associated in the first place. This creates a problem of restricted range: Model performance is only assessed in a very narrow range of the plausible variation of word association strengths.

We believe it is prudent to discount the magnitude of the correlation observed between measures of forward association strength and differences along particular PCs. The more

informative result is that PCs extracted early predict forward associative strength better. This suggests that earlier PCs matter most to human organization of semantic knowledge.

Prediction of lexical access

Another way to approach assessing the psychological plausibility of skip-gram PCs is by using them to predict measures of lexical access (e.g., lexical decision, word naming), a process that requires humans to access semantic knowledge. We examined the ability of skip-gram PCs to predict three measures of lexical access: lexical decision times (LD-RTs) and word-naming times (NMG-RTs) taken from the English Lexicon Project (ELP; Balota et al., 2007), and LD-RTs taken from the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012).

In all, 70 PCs reliably predicted LD-RTs within the ELP data, 58 PCs reliably predicted NMG-RTs within the ELP data, and 63 PCs reliably predicted LD-RTs within the BLP data. Relative to the lexical and affective measures, behavioral measures have relationships to a broader range of PCs. This is expected, since multiple semantic considerations bear on behavioral tasks. Of the 63 PCs that reliably predicted BLP LD-RTs, only 43 (68 %) also reliably predicted ELP LD-RTs. In contrast, 53 (83 %) of the PCs that reliably predicted ELP NMG-RTs also predicted ELP LD-RTs. It would seem that the semantic considerations affecting lexical decision data from the ELP are more like those affecting word-naming data from that corpus than like the semantic considerations affecting lexical decision data from the BLP.

Of all the dimensions, PCs 3 and 4 have the strongest relationships to both ELP LD-RTs [PC 3, $r(12{,}687) = .14$, $p < 2.2e{-}16$; PC 4, $r(12{,}687) = .21$, $p < 2.2e{-}16$] and ELP NMG-RTs [PC 3, $r(12{,}687) = .15$, $p < 2.2e{-}16$; PC 4, $r(12{,}687) = .26$; $p < 2.2e{-}16$]. PC4 was labeled a dimension of *meaning specificity*, on the basis of previous analyses. PC3 remained unlabeled, but was likewise correlated with our measure of *meaning specificity*. In contrast, only PC1 had a relationship strength of similar magnitude to BLP LD-RTs [$r(7{,}817) = .21$, $p < 2.2e{-}16$]. PC1 was labeled a dimension of *word frequency*, on the basis of previous analyses.

Overall, LD-RTs are related to a broader range of PCs than are NMG-RTs. This is perhaps because lexical decision requires more analysis of the word strings than word naming does, as reflected in the fact that the average ELP NMG-RT values (Average [*SD*] = 679 [80] ms) are reliably shorter than the average ELP LD-RT values (Average [*SD*] = 724 [103] ms) [$t(12{,}688) = 66.8$, $p < 2.2e{-}16$].

We regressed skip-gram PCs on lexical access times after splitting our available data into two equal halves: a training set and a validation set. PCs were selected for inclusion on the basis of having a statistically reliable correlation with the outcome variable ($a = .05/300$). Since our predictors are orthogonal by definition, we can assume that if a predictor had a statistically reliable relationship with the outcome variable when it was used as the sole predictor, it would maintain its predictive validity when included in a multiple regression involving other PCs. Models were constructed on their training set and later tested on their validation set.

The regression models accounted for 30.49 % of the variance in ELP LD-RTs (cross validation: 28.76 %), 24.18 % of the variance in NMG-RTs (cross validation: 24.59 %), and 34.63 % of the variance in BLP LD-RTs (cross validation: 33.24 %). We refer to the overall contributions of the skip-gram predictors to the behavioral measures of lexical access as *sgLD-E*, *sgNMG-E*, and *sgLD-B* for the ELP lexical decision and naming data and the BLP lexical decision data, respectively. We estimated these values by rerunning each regression on the combined validation and test sets and taking the resulting model's estimates of response times.

Most models of lexical access do not account for a large proportion of variance that could theoretically be accounted for in behavioral measures (Adelman, Marquis, Sabatos-DeVito, & Estes, 2013). Adelman et al. (2013) suggested that some of this variance could be due to semantic effects, which are currently underrepresented in models of lexical access. This has spurred an examination of affective factors (Kuperman et al., 2014). Our analysis suggests that some of this missing variance may be illuminated by the semantic space of skip-gram PCs.

A useful benchmark for testing the predictive validity of lexical access models is the three-term model of log frequency, orthographic neighborhood size, and word length. Across our data, this three-term model accounted for 45.1 % of the variance in ELP LD-RTs, 36.9 % of the variance in ELP NMG-RTs, and 41.46 % of the variance in BLP LD-RTs.

When sgLD-E or sgNMG-E were added as predictors, respectively, 48.58 % of the variance in ELP LD-RTs and 40.51 % of the variance in ELP NMG-RTs was explained. When sgLD-B was added as a predictor, 46.82 % of the variance in BLP LD-RTs was explained. When the three-term base models were compared to the four-term base + skip-gram models, the base + skip-gram models provided superior fits in all three cases (ELP LD-RT, $F(1, 12687) = 858.06$, $p < 2.2e{-}16$; ELP NMG-RT, $F(1, 12687) = 770.84$, $p < 2.2e{-}16$; BLP LD-RT, $F(1, 7633) = 770.23$, $p < 2.2e{-}16$).

It is possible that this increase in fit was due to the fact that the skip-gram model captured aspects of the semantic variables we identified earlier as correlated with PCs, all of which are known to influence lexical access. To assess this possibility, models that also included valence, arousal, and dominance judgments (Warriner et al., 2013), concreteness judgments (Brysbaert et al., 2014), and age-of-acquisition judgments (Kuperman et al., 2012) were considered.

We compared two additional classes of models. First we assessed the three-term base model plus valence, arousal,

dominance, concreteness, and age of acquisition. We also include interactions between all of the terms and log frequency, since lexical access effects typically are mediated by word frequency (but see Baayen, 2010). The second class of models contained these terms, plus sgLD-E, sgNMG-E, or sgLD-B. The 15-term base + semantic + interaction models accounted for 53.47 % of the variance in ELP LD-RTs, 46.10 % of the variance in ELP NMG-RTs, and 50.98 % of the variance in BLP LD-RTs.

Adding in an additional predictor derived from skip-gram PCs resulted in models that accounted for 54.30 % of the variance in ELP LD-RTs, 46.96 % of the variance in ELP NMG-RTs, and 53.39 % of the variance in BLP LD-RTs. In all three cases, the models including the skip-gram terms provided superior fits [ELP LD, $F(1, 12318) = 224.15$, $p < 2.2e–16$; ELP NMG, $F(1, 12318) = 198.42$, $p < 2.2e–16$; BLP LD, $F(1, 7397) = 382.64$, $p < 2.2e–16$]. Semantic dimensions derived from skip-gram PCs accounted for unique variance in behavioral measures of lexical access that was not accounted for by any of our eight lexical and semantic predictors, nor by their interactions with log frequency.

The performance of the five reported models on the three measures of lexical access is provided in Table 2.

Potential semantic variables

We attempted to identify "new" semantic variables that might substantially impinge on lexical access, beyond the five considered throughout our present work. For each of our measures of lexical access, forward stepwise regression was conducted for five steps using the PCs that had previously been identified as being reliably related to that measure of lexical access. At each forward step, terms were selected for inclusion if they resulted in any reduction of information loss, based on the Akaike information criterion with a $k = 2$ penalty for extra parameters. We looked for terms that entered into the regressions for two or more of the measures of lexical access. We used a limitation of five forward steps to restrict ourselves to only those PCs most strongly related to measures of lexical access.

Two variables entered into all three of the models: PC18 and PC35. Additionally, PC3 and PC24 entered into the models of ELP LD-RT and ELP NMG-RT. Readers are reminded that PC3 was earlier suggested to be a dimension of agency. Correlations between these PCs and our three measures of lexical access can be found in Table 3. In all cases, highly reliable effects were observed ($p < 2.2e–16$).

We were unable to identify any organizing theme by examining words that loaded high or low on PC18. High-loading words included *aura*, *excitement*, *copycat*, *spotlight*, and *exposure*. Low-loading words included *march*, *outboard*, *axle*, *gallant*, and *monsieur*.

One pole of PC35 had a tendency toward including words having to do with threat and/or a medical setting. The five highest-loading words were *pricey*, *preemptive*, *doomsday*, *colonoscopy*, and *layaway*. The 20 highest-loading words also included words like *invasive*, *surgery*, *beeper*, *heresy*, *hysterectomy*, *appendectomy*, *neurosurgery*, *smallpox*, and *pacemaker*. The other pole contained words noisily tied to gender/sexuality—for example, *cadet*, *gross*, *barmaid*, *transvestite*, *schoolgirl*, *girl*, *lass*, and *sixteen*.

The words at either pole of PC24 clustered together in meaning, but it was unclear how they would be related to some, more continuous dimension of meaning. High-loading words were related to affiliation and/or criminality. They included *oppressed*, *brown*, *gritty*, *blue*, *gang*, *uniform*, *militia*, *scrawny*, *blank*, *dusty*, *illiterate*, *bearded*, *bandana*, *fingerprint*, *armed*, *black*, *thumbprint*, *nameless*, *white*, and *uniformed*. These words include actual groups (*gang*, *militia*), symbols of affiliation (*brown*, *blue*, *black*, *white*, *uniform*, *bandana*), identifying descriptors (*scrawny*, *bearded*), among other words related to social power dynamics (*oppressed*, *illiterate*, *fingerprint*). Low-loading words were related to

Table 2 Comparison of five models in terms of the amounts of variance accounted for in three measures of lexical access: Lexical decision and word naming times from the English Lexicon Project (ELP; Balota et al., 2007), and lexical decision times from the British Lexicon Project (BLP; Keuleers et al., 2012)

	Predictor Sets				
	skipgram	Lexical	Lexical + skipgram	Lexical + Semantic	Lexical + Semantic + skipgram
Lexical decision (ELP)	.288	.451	.486	.538	.543
Word naming (ELP)	.246	.369	.410	.461	.470
Lexical decision (BLP)	.332	.414	.468	.510	.534

The models included up to three predictor sets: the lexical variables of log frequency, word length, and orthographic neighborhood size (lexical); the semantic variables of valence, arousal, dominance, concreteness, and age of acquisition (semantic); and a single variable derived from skip-gram principal components. In the event that both semantic and lexical variables were entered into the same model, all semantic interactions with log frequency were also included. In all cases, adding the skip-gram predictor to a model reliably increased the proportion of variance that model accounted for ($p < 2.2e–16$)

Table 3 Correlation strengths between Principal Components 3, 18, 24, and 35 and three behavioral measures of lexical access

	Principal Component			
Variable	PC3	PC18	PC24	PC35
Lexical decision (ELP)	.14	.10	.13	.11
Naming times (ELP)	.15	.08	.12	.08
Lexical decision (BLP)	.07	.12	.12	.11

Of the PCs not provided a semantic label, these four PCs were identified as those most strongly related to measures of lexical access

high positive valence, particularly high-valence words related to the presence of water. They included *extravagant*, *extravagance*, *stunt*, *lavish*, *aquarium*, *yacht*, *attraction*, *spectacle*, *diver*, *outrageous*, *scandalous*, *barge*, *lagoon*, *rainmaker*, *gaff*, *speedboat*, *boat*, *spectacular*, *whirlpool*, and *caterer*.

We are uncertain what to conclude about this examination of PC18, PC24, and PC35, other than pointing to the fact (Table 3) that they account for amounts of variance in the lexical access measures comparable to those seen for more the established semantic variables valence, arousal, dominance (Warriner et al., 2013), and concreteness (Brysbaert et al., 2014). We believe a more thorough examination of these PCs is worth further investigation.

Taken together, the results of this section suggest that skip-gram PCs are picking up on psychologically relevant aspects of semantic judgments. Furthermore, they are picking up on information relevant to lexical access not carried by typical lexical and semantic variables. Direct study of how the skip-gram model organizes word meaning may provide insights to the organization of human lexical semantics.

Discussion

Motivated by the observations that the skip-gram model makes meaningful semantic judgments and that it may be representing meaning within a linear, high-dimensional space, we attempted to identify those dimensions.

Our analysis of the PCs extracted from skip-gram vectors provided evidence that the skip-gram model organizes meaning in a way that maps onto lexical and semantic concepts that are standard in psycholinguistic research. This is an interesting finding, because the skip-gram model was never explicitly supplied with such information. Rather, the representation of these concepts was the consequence of learning to identify sources of variation that allow the skip-gram model to map words to their contexts of occurrence. This does not necessarily mean that the skip-gram model is a psychologically plausible model of semantics. The semantic dimensions we extracted may be a consequence of the text from which the skip-gram model learned, not the learning algorithm itself.

The work of Osgood and his colleagues (Osgood et al., 1957) converges on the idea that humans largely make semantic distinctions along affective dimensions. Our results support and extend this conclusion. We found that the PCs extracted earliest from the skip-gram model have the highest predictive validity for predicting forward association strength. Of the seven earliest PCs, two (PC5, PC7) were identified as being related to affective human judgments of *valence* and *dominance*. These variables, in turn, are tied closely to Osgood's affective concepts of *evaluation* and *potency*, respectively.

Recent research has suggested that *arousal* (a construct similar to Osgood's *activity*) also impinges on lexical semantics (e.g., Kuperman, Estes, Brysbaert, & Warriner, 2014). Although we found that numerous PCs are reliably correlated with human judgments of arousal, we did not find any particular PC that met objective criteria for being labeled an "arousal dimension." Rather, the one dimension that most strongly discriminated according to arousal (PC2) was more plausibly interpreted as a dimension of concreteness.

Relative to other affective measures, measures of arousal are not particularly reliable: Warriner et al. (2013) reported split-half reliabilities of .69 within their 13,000-word norm set (as compared to .91 for valence and .77 for dominance). When comparing between demographic groups defined by male versus female, young versus old, and high versus low education, the correlations between arousal judgments were only .52, .50, and .41, respectively (as compared to .79, .82, and .83 for valence, and .59, .59, and .61 for dominance). Arousal judgments from the ANEW norms set (Bradley & Lang, 1999) correlated with those in the Warriner et al. (2013) norms set at $r = .76$ (as compared to .95 for valence and .80 for dominance). Attempts at algorithmically extrapolating human judgments have likewise found that human judgments of arousal are much less predictable than other semantic measures (Hollis & Westbury, 2016; Mandera, Keuleers, & Brysbaert, 2015; Recchia & Louwerse, 2015; Westbury et al., 2015). These findings, along with the results of our analysis of the skip-gram model, suggest that the concept of arousal is not as clearly specified as a semantic construct. Our results suggest that further understanding of the construct of arousal may be made by interpreting its relationship to concreteness.

Westbury et al. (2013) demonstrated that affective measures can account for a large portion of the psychologically relevant variation contained within human judgments of imageability. Computer-estimated imageability judgments (from Westbury et al., 2013), in turn, are closely related to Brysbaert, Warriner, and Kuperman's (2014) concreteness judgments ($r = .77$ over 5,278 words). The explanation we propose for why judgments of concreteness are so strongly related to PC2 is that, like imageability judgments, concreteness judgments are affective in nature. In addition to being correlated with arousal ($r = .27$), PC2 is correlated with valence ($r = .21$) and dominance ($r = .14$). This groups PC2

along with PC5 and PC7 as another affective dimension along which the skip-gram model organizes word meaning.

One of our more surprising findings was that information carried by lexical variables (log frequency and word length) can be partially reconstructed from a PCA of skip-gram vectors. This is surprising because lexical features are never presented to the skip-gram model in training, and it is trained to make a semantic discrimination, not a lexical discrimination. The conclusion that we draw from this is that log frequency and word length are not exclusively lexical properties: They carry semantic information, too, as Harris (1970) pointed out when he wrote "[i]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with differences of distribution" (p. 13).

This is important, since lexical effects are typically assumed to be informative of the underlying functional architecture of our semantic system, not about the organization of semantics (though see Franklin & Mewhort, 2015, and Jones & Mewhort, 2007, for a challenge to those assumptions). Our findings are consistent with arguments that word frequency has little psychological validity (e.g., Adelman, Brown, & Quesada, 2006; Baayen, 2010; McDonald & Shillcock, 2001), suggesting rather that the apparently ubiquitous effects of word frequency on lexical access are symptomatic of it being intertwined with informational and semantic considerations.

The claim that word frequency is not an ontologically primary concept helps reconcile the contradiction that, on the one hand, affect is a relevant aspect of semantics (Osgood et al., 1957) but, on the other hand, affective measures account for only a small portion of the variance in behavioral measures of lexical access (Kuperman et al., 2014). Frequency effects may simply be side-effects of the fact that word meaning organizes word use, and many semantic effects may be obscured by word frequency effects. The inspection of words with high and load loadings on PC1 suggests that two possible semantic considerations organizing variation in word frequency are Osgood et al.'s affective concept of *dynamism* and/or knowledge pertaining to *gender*.

Our analysis suggests that the skip-gram model partially reconstructs word length information because word length carries information about meaning specificity, and meaning specificity is a useful dimension along which to organize word meaning (PC4). As with frequency, this claim suggests that the effects attributed to word length may instead index contributors to lexical access that have nothing to do with word length per se.

Both PC1 and PC4 (the PCs most strongly related to frequency and word length, respectively) were also moderately correlated with human judgments of concreteness ($r = .39$, $r = .28$, respectively). If, as we suggest above, judgments of concreteness are affective in nature, this also implicates PC1 and PC4 as possible affective dimensions, tying together all five of our labeled PCs (PC1 = log frequency, PC2 = concreteness, PC4 = meaning specificity, PC5 = valence, PC7 = dominance), to varying degrees, as affective dimensions.

The last main contribution of this work is a demonstration that the PCs extracted from the skip-gram model account for variance in lexical access measures that cannot be accounted for by traditional lexical and semantic measures recognized by psycholinguists. A more detailed study of how the skip-gram model organizes semantic concepts may thus provide illumination regarding how humans do the same. We identified four PCs that consistently accounted for variation across lexical access tasks but were not strongly related to one of our eight lexical or semantic variables: PC3, PC18, PC24, and PC35. Although our analysis of the organizing themes of these four PCs was not conclusive, we point out that they are about as predictive of behavioral measures of lexical access as the affective constructs that up to now have received attention in the study of lexical access (Estes & Adelman, 2008a, 2008b; Kuperman et al., 2014; Larsen, Mercer, Balota, & Strube, 2008). We believe that a more thorough content analysis of the skip-gram model's semantic space, and of these four PCs in particular, will shed light on the organization of human lexical semantics.

Author note We thank Marc Brysbaert and two anonymous reviewers for helpful advice on an earlier draft of the manuscript.

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17,* 814–823. doi:10.1111/j.1467-9280.2006.01787.x

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

Andonova, E., D'Amico, S., Devescovi, A., & Bates, E. (2004). Gender and lexical access in Bulgarian. *Perception & Psychophysics, 66,* 496–507.

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

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

Baron, N. S. (1971). A reanalysis of English grammatical gender. *Lingua, 27,* 113–140.

Bender, A., Beller, S., & Klauer, K. C. (2016). Crossing grammar and biology for gender categorisations: Investigating the gender congruency effect in generic nouns for animates. *Journal of Cognitive Psychology*. doi:10.1080/20445911.2016.1148042. **Advance online publication**.

Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in*

mind: Advances in the study of language and thought (pp. 61–79). Cambridge, MA: MIT Press.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings (Technical Report C-1)*. Gainesville, FL: University of Florida, NIMH Center for Research in Psychophysiology.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46,* 904–911. doi:10.3758/s13428-013-0403-5

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

Esaulova, Y., & Von Stockhausen, L. (2015). Cross-linguistic evidence for gender as a prominence feature. *Frontiers in Psychology, 6,* 1356. doi:10.3389/fpsyg.2010.00174

Estes, Z., & Adelman, J. S. (2008a). Automatic vigilance for negative words in lexical decision and naming: Comment on Larsen, Mercer, and Balota (2006). *Emotion, 8,* 441–444. doi:10.1037/1528-3542.8.4.441

Estes, Z., & Adelman, J. S. (2008b). Automatic vigilance for negative words is categorical and general. *Emotion, 8,* 453–457. doi:10.1037/a0012887

Fang, S., Murumatsu, K., & Matsui, T. (2015). Experimental study of aesthetic evaluation to multi-color stimuli using semantic differential method. *Transactions of Japan Society of Kansei Engineering, 14,* 37–47.

Franklin, D. R. J., & Mewhort, D. J. K. (2015). Memory as a hologram: An analysis of learning and recall. *Canadian Journal of Experimental Psychology, 69,* 115–135. doi:10.1037/cep0000035

Fredrickson, B. L., & Roberts, T. A. (1997). Objectification theory. *Psychology of Women Quarterly, 21,* 173–206.

Harris, Z. (1970). *Papers on syntax* (H. Hiz, Ed.). Boston, MA: Reidel.

Hollis, G., & Westbury, C. F. Extrapolating Human Judgments from Word2Vec Vector Representations of Word Meaning. *The Quarterly Journal of Experimental Psychology*, (2016).

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

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44,* 287–304. doi:10.3758/s13428-011-0118-4

Konishi, T. (1993). The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research, 22,* 519–534.

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General, 143,* 1065–1081. doi:10.1037/a0035669

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44,* 978–990. doi:10.3758/s13428-012-0210-4

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240. doi:10.1037/0033-295X.104.2.211

Larsen, R. J., Mercer, K. A., Balota, D. A., & Strube, M. J. (2008). Not all negative words slow down lexical decision and naming speed: Importance of word arousal. *Emotion, 8,* 445–452. doi:10.1037/1528-3542.8.4.445

Levy, O., & Goldberg, Y. (2014a). Linguistic regularities in sparse and explicit word representations. In R. Morante & S. W. Yih (Eds.), *Proceedings of the Eighteenth Conference on Computational Natural Language Learning* (pp. 171–180). Stroudsburg, PA: Association for Computational Linguistic.

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

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

MacKay, D. G. (1999). Gender in English, German, and other languages: Problems with the old theory, opportunities for the new. In U. Pasero & F. Braun (Eds.), *Herstellung und Wahrnehmung von Geschlech [Perceiving and performing gender]* (pp. 73–87). Viesbaden, Germany: Westdeutscher Verlag.

Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, & Computers, 36,* 421–431. doi:10.3758/BF03195590

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *Quarterly Journal of Experimental Psychology, 68,* 1623–1642.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech, 44,* 295–322.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology, 14,* 261–292.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from 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 26* (pp. 3111–3119). Cambridge, MA: MIT Press.

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Human Language Technologies—North American Association for Computational Linguistics* (pp. 746–751). Stroudsburg, PA: Association for Computational Linguistics.

Mills, A. E. (1986). Acquisition of the natural-gender rule in English and German. *Linguistics, 24,* 31–46.

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems 26* (pp. 2265–2273). Cambridge, MA: MIT Press.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36,* 402–407. doi:10.3758/BF03195588

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

Ou, L. C., Luo, M. R., Woodcock, A., & Wright, A. (2004a). A study of colour emotion and colour preference. Part I: Colour emotions for single colours. *Color Research and Application, 29,* 232–240.

Ou, L. C., Luo, M. R., Woodcock, A., & Wright, A. (2004b). A study of colour emotion and colour preference. Part II: Colour emotions for two-colour combinations. *Color Research and Application, 29,* 292–298.

Qiu, L., Cao, Y., Nie, Z., Yu, Y., & Rui, Y. (2014). Learning word representation considering proximity and ambiguity. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 1572–1578). Washington, DC: AAAI Press.

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *Quarterly Journal of Experimental Psychology, 68,* 1584–1598.

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

Scheutz, M. J., & Eberhard, K. M. (2004). Effects of morphosyntactic gender features in bilingual language processing. *Cognitive Science, 28,* 559–588.

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

Twain, M. (1880). *A tramp abroad*. Leipzig, Germany: Tauchnitz.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45,* 1191–1207. doi:10.3758/s13428-012-0314-x

Westbury, C. (2014). You Can't Drink a Word: Lexical and Individual Emotionality Affect Subjective *Familiarity Judgments. Journal of psycholinguistic research, 43*, 631–649.

Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2015). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *Quarterly Journal of Experimental Psychology, 68,* 1599–1622. doi:10.1080/17470218.2014.970204

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: On emotion, context, and the algorithmic prediction of human imageability judgments. *Frontiers in Psychology, 4,* 991. doi:10.3389/fpsyg.2013.00991

Whorf, B. L. (1945). Grammatical categories. *Language, 21,* 1–11.

Word2vec. (2013). word2vec: Tool for computing continuous distributed representations of words. Retrieved October 18, 2015, from https://code.google.com/p/word2vec/