Cognitive Psychology 131 (2021) 101441

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Distributional social semantics: Inferring word meanings from communication patterns

Brendan T. Johns

Department of Psychology, McGill University, 2001 McGill College Avenue, Montreal, Quebec H3A 1G1, Canada

ARTICLE INFO

Keywords:
Lexical semantics
Distributional modeling
Cognitive modeling
Machine learning
Big data

ABSTRACT

Distributional models of lexical semantics have proven to be powerful accounts of how word meanings are acquired from the natural language environment (Günther, Rinaldi, & Marelli, 2019; Kumar, 2020). Standard models of this type acquire the meaning of words through the learning of word co-occurrence statistics across large corpora. However, these models ignore social and communicative aspects of language processing, which is considered central to usage-based and adaptive theories of language (Tomasello, 2003; Beckner et al., 2009). Johns (2021) recently demonstrated that integrating social and communicative information into a lexical strength measure allowed for benchmark fits to be attained for lexical organization data, indicating that the social world contains important statistical information for language learning and processing. Through the analysis of the communication patterns of over 330,000 individuals on the online forum Reddit, totaling approximately 55 billion words of text, the findings of the current article demonstrates that social information about word usage allows for unique aspects of a word's meaning to be acquired, providing a new pathway for distributional model development.

1. Introduction

The natural language environment is statistically redundant, with word usage being consistent across many occurrences of a given word. This consistency allows for sophisticated representations of word meanings to be acquired when simple learning mechanisms are used to learn from large samples of natural language. This class of model, entitled distributional models of semantics, comes in a variety of forms (e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013; for reviews, see Günther, Rinaldi, & Marelli, 2019; Jones, Willits, & Dennis, 2014; Kumar, 2020). Distributional models are based on the premise that how a word is used across a sizeable subsection of natural language provides evidence about the meaning of that word. Firth (1957) expressed the intuition behind these models by stating "You shall know a word by the company it keeps." The end result of the distributional modeling process are vector representations of word meanings, which have been shown to provide a good account of a variety of lexical semantic data.

The first major model in this line of research is the classic Latent Semantic Analysis (LSA) model of Landauer and Dumais (1997). This work set the stage for the current machine learning and big data approaches in the cognitive and language sciences by demonstrating the systematic connection between the structure of the language environment and internalized representations of knowledge, a classic problem in the philosophy of mind dating back to Plato. LSA established that a large-scale examination of the

E-mail address: brendan.johns@mcgill.ca.

https://doi.org/10.1016/j.cogpsych.2021.101441
Received 20 January 2021; Received in revised form 5 October 2021; Accepted 7 October 2021
Available online 16 October 2021
0010-0285/© 2021 Elsevier Inc. All rights reserved.

natural language environment allowed for accurate representations of word meanings to be extracted from realistic linguistic experience.

The progress that has been made in this domain since Landauer and Dumais (1997) is considerable. Indeed, distributional models are now a key model type in the cognitive, psychological, and language sciences (Johns, Jamieson, & Jones, 2020), and have been used in computational modeling across a number of domains, including episodic memory (e.g., Johns, Jones, & Mewhort, 2012, 2021; Mewhort, Shabahang, & Franklin, 2018; Osth, Shabahang, Mewhort, & Heathcote, 2020), lexical organization (e.g., Hoffman, Ralph, & Rogers, 2013; Hollis, 2017; Hsiao & Nation, 2018; Jones, Johns, & Recchia, 2012; Johns, Dye, & Jones, 2020; Johns, 2021), morphological processing (e.g., Marelli & Baroni, 2015; Marelli, Gagné, & Spalding, 2017; Westbury & Hollis, 2019), verbal fluency (e.g., Hills, Jones, & Todd, 2012; Johns et al., 2018; Taler, Johns, Young, Sheppard, & Jones, 2013; Taler, Johns, & Jones, 2020), decision (e.g., Bhatia, 2017; Bhatia & Stewart, 2018; Bhatia, Richie, & Zou, 2019), and sentence processing (e.g., Johns & Jones, 2015; Johns, Jamieson, Crump, Jones, & Mewhort, 2020), among others.

A current trend in distributional semantics in both cognitive and machine learning research is the integration of perceptual information into the lexical representations of words (e.g., Bruni, Tran, & Tran, 2014; De Deyne, Navarro, Collell, & Perfors, 2021; Lazaridou, Marelli, & Baroni, 2017), in recognition of the grounding problem that distributional semantic models face (Andrews, Vigliocco, & Vinson, 2009; Johns & Jones, 2012; Riordan & Jones, 2011). In these approaches, linguistic representations of word meanings derived from large corpora have been combined with multi-modal perceptual information, in order to ground their representations in the perceptual world (such as those derived from feature and grounded production norms; e.g., Lynott, Connell, Brysbaert, Brand, & Carney, 2019; McRae, Cree, Seidenberg, & McNorgan, 2005), due to the continued recognition of the importance of perceptually grounded information in language processing (e.g., Barsalou, 1999, 2008).

However, there are other extra-linguistic types of information that standard distributional models ignore. In particular, they do not include social information about language usage, even though this information type is considered central to developmental theories of language, such as in usage-based (Tomasello, 2003, 2009) and adaptive (Beckner et al., 2009; Christiansen & Chater, 2008) approaches. These theories are based on the idea that language is a fundamentally communicative and a social tool and is acquired through the examination of how others use language. As an example of the findings in this line of research, Lieven, Pine, and Baldwin (1997) demonstrated that the majority of a child's utterances are based on a few recently experienced lexical patterns from others in their environment. Computational models based on usage-based theories is a growing trend (see Bannard, Lieven, & Tomasello, 2009; Johns & Jones, 2015; Johns, Jamieson, Crump, Jones, & Mewhort, 2020; see also Abbot-Smith & Tomasello, 2006 for an earlier proposal), suggesting that the mechanisms put forth in these theories are credible targets for continued computational model development.

Adaptive and usage-based theories of language propose that an efficient way to learn a language is to communicate as others do within their environment. That is, given some personal need (e.g., I would like some water), if one experiences another person satisfy that need through communication (e.g., seeing someone ask "May I have a glass of water?"), that experience provides a tremendous amount of information about how to communicate one's wants and needs without the need to form sophisticated abstractions. Scaled across thousands or millions of episodic experiences with language (Brysbaert, Stevens, Mandera, & Keuleers, 2016 estimate that a typical student should experience more than ten-million word tokens through spoken language every year), these experiences provide the basis for acquiring sophisticated representations of language which are likely not based in abstracted grammatical rules (Tomasello, 2003). Since the meanings of words in a language are intimately tied to their communicative functions (Borghi and Binkofski, 2014; Gärdenfors, 2018), it is probable that communicative patterns of word usage should provide information about the meaning of words.

A recent example of the power of combining large-scale computational cognitive modeling with usage-based and adaptive theories of language, Johns, Jamieson, Crump, Jones, and Mewhort (2020; see Jamieson, Avery, Johns, & Jones, 2018 and Johns & Jones, 2015 for earlier work) proposed an instance-based distributional model of sentence production and syntactic priming (Bock, 1986; for a review, see Pickering & Ferreira, 2008). This model stored instances of sentences (based on the BEAGLE model of Jones & Mewhort, 2007), and produced syntactically correct sentences based on the latent patterns of word order contained across those instances (using the retrieval operations of the classic episodic memory model MINERVA 2; Hintzman, 1986). Syntactic priming was accounted for in the model by assuming that an adaptive mechanism for language production is to communicate like others in your environment are producing language. Thus, a syntactic prime is a signal of how to communicate effectively, as it indicates how others are sharing information in context. When the model was biased with a prime sentence, the model tended to prefer the same syntactic construction as the prime used, even when there was no semantic overlap between prime and target sentences. This result shows the promise of integrating distributional technologies together with socially-based linguistic theories to develop new solutions to old problems.

As a first demonstration of the power of utilizing large-scale social information to account for language processing, Johns (2021) recently proposed a distributional model of lexical strength based around text corpora derived from the online discussion forum Reddit, attained from Baumgartner, Zannettou, Keegan, Squire, and Blackburn (2020), in order to develop a communicative account of lexical organization. The lexical strength measures derived from these materials were based on the communication patterns of words across users and discourses. The ability to extract discourse information comes from 'subreddits' which contain submissions and comments surrounding a given topic (e.g., r/cogsci contains discussions centered around cognitive science). The ability to extract user information comes from the recording of comments that individual users (who had publicly available usernames) produced across subreddits.

The models developed by Johns (2021) were based around the contextual diversity proposals of Adelman, Brown, & Quesada (2006), who theorized that instead of counting the frequency of words (as classic models of lexical organization propose; e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Goldinger, 1998; Murray & Forster, 2004), one should instead count the number of

contexts that a word occurs in, in order to best account for lexical organization data (see Jones, Dye, & Johns, 2017 for a review). Johns (2021) modified this proposal by defining linguistic context as either an individual user or a discourse, which is a much larger notion of linguistic context than was previously considered (see also Hollis, 2020; Johns, Dye, & Jones, 2020; Johns & Jones, 2021 for related work on how to define linguistic contexts). These count models were modified by a distributional model entitled the Semantic Distinctiveness Model (SDM; Jones, Johns, & Recchia, 2012), which is a distributional model of lexical strength that provides a graded strength to a context depending on how unique that context is compared to a word's event history. It was found that the socially-based models developed provided benchmark fits to a variety of different large-scale lexical organization data, suggesting that social and communicative information about language usage provides important information about how language is organized in the mind. Additionally, Johns (in press) demonstrated that these findings generalize to item-level variability in episodic recognition, demonstrating the universality of the derived variables.

The results of Johns (2021) demonstrate that integrating social information into a distributional model provides additional power to a model of lexical organization. There are a wide variety of results demonstrating the impact of semantics on lexical organization (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Buchanan, Westbury, & Burgess, 2001; Hoffman et al., 2013; Jones, Johns, & Recchia, 2012; Yap, Tan, Pexman, & Hargreaves, 2011), suggesting that semantics and lexical organization are intimately connected. The goal of this article is to determine if communication patterns of word usage, embedded within a distributional modeling framework, can account for unique variance in word similarity and relatedness data, similar to what was found for lexical organization in Johns (2021). Such a finding would demonstrate the importance of communicative information to lexical semantic learning and processing.

The role of social and communicative information to lexical semantics will be answered through the analysis of online communication patterns on the discussion forum Reddit. The proposed models will be trained on the distributional communication patterns of over 330,000 individuals across 30,000 discourses (totaling approximately 55 billion words of written language). The representations that the socially-based models construct will be contrasted with a standard distributional model which learns from word-word co-occurrence statistics, in order to determine if the communication patterns of words provide a unique signal about word meanings.

The underlying theoretical motivation of this research is the recognition of the importance of the social world in language comprehension and processing, based on usage-based and adaptive theories of language (Beckner et al., 2009; Christiansen & Chater, 2008; Tomasello, 2003, 2009). These theories entail that for distributional models of language and memory to continue to be developed as plausible models of human cognition, they will need to integrate social information into their processing framework. The work described here is a first attempt to determine if social information plays an important role in the computation of word similarity and acquisition of word meanings. The results of this article will provide an existence proof that social information about language usage does provide unique information about word meanings. The first section of the paper will describe the distributional modeling framework and the various proposed models, followed by analyses designed to determine if models based in social and communicative information provide a unique information source about lexical semantics and lexical organization.

2. Modeling framework

There are a number of different learning mechanisms that are employed in distributional modeling, such as matrix decomposition (Griffiths, Steyvers, & Tenenbaum, 2007; Landauer & Dumais, 1997), vector accumulation (Jones & Mewhort, 2007), neural embedding (Mikolov et al., 2013), and retrieval-based models (Jamieson, Avery, Johns, & Jones, 2018), among others. However, the central component of all models is the exploitation of word-word co-occurrence patterns. That is, words that co-occur together across linguistic contexts tend to have more similar semantic representations at the end of model training. Indeed, an information theoretic measure of the likelihood of two words co-occurring together in the same context, entitled pointwise mutual information (PMI), has been shown to provide an excellent account of word similarity data (Bullinaria & Levy, 2007, 2012; Levy & Goldberg, 2014; Levy, Goldberg, & Dagan, 2015; Recchia & Jones, 2009). This model type is typically referred to as a count-based model, as it simply counts the co-occurrence patterns of words across contexts.

The modeling framework used here is similar to past count-based models, combined with some of the training methodologies used by neural embedding models (Mikolov et al., 2013), recently proposed by Johns, Mewhort, and Jones (2019). The underlying representation used in the model is a matrix, where each row contains a word's representation, and each column is a feature of some type. In Johns et al. (2019), the main feature type used was word co-occurrence with other words within a window in a sentence. That is, each word's semantic representation is the number of times that it occurred with other words in a sentential context. The resulting matrix for this model has a dimensionality of $W \times W$, where W is the number of words in the model's vocabulary, and each entry in the matrix is a count of the number of times two words co-occurred with each other in context. This model will be referred to as the WW model.

To construct an accurate semantic representation with the WW model, two transformations were applied to the co-occurrence matrix, informed by the training mechanisms used by neural embedding models. In neural embedding models (e.g., word2vec; Baroni, Dinu, & Kruszewski, 2014; Mikolov et al., 2013), neural networks are trained to predict the words that should co-occur with a target word in a sentential context. Incorrect predictions provide an error signal used to bring the hidden weights of the network closer to the base rate of environmental co-occurrence of a training corpus, modifying the network with backpropagation (among other optimization techniques). This is referred to as the learning of positive information. Neural embedding models employ an additional training mechanism, entitled negative sampling. In negative sampling, a set number of random words are sampled (based upon word frequency, with more frequent words being more likely to be sampled). The network is then trained to suppress those randomly sampled words in the predicted output layer. The idea is that the network should be able to both predict words that a word should co-

occur with, as well as those words that they do not co-occur with. Negative sampling has repeatedly been shown to have a sizeable effect on the model's performance (e.g., Goldberg & Levy, 2014; Levy et al., 2015).

The goal of Johns et al. (2019) was to determine if the importance of negative sampling was due to the honing of a prediction mechanism (as initially proposed in the development of word2vec), or whether the practice offers some additional information about word meanings not captured by previous distributional modeling frameworks. The results of this study indicated that the utility of negative sampling is not in the honing of a prediction mechanism, which was found by demonstrating that the WW model showed the same advantage for negative sampling as neural network models do. Given that the WW does not use prediction in forming a semantic representation, the advantage that this technique provides is not due to improving the predictive power of a neural network. Instead, it was shown that negative sampling serves to integrate base-rate occurrence of words into a word's semantic representation. The integration of this information allows for the unique associations between words to be highlighted. If the co-occurrence rate between two words vastly exceeds the base rate of occurrence of those words, it signals that those two words have a strong association to each other. If the co-occurrence rate between those two words does not exceed a base-rate, then the association between those two words is not likely to be informative about the meaning of that word. Johns et al. (2019) demonstrated that the integration of a negative sampling routine vastly increases the performance of the WW model. Additionally, they described two parameter-free analytical transformations of a matrix that allows for negative information to be readily integrated into a model's lexical representation, eliminating the need for a sampling procedure.

The application of these techniques to the WW matrix allowed for the resulting model to achieve similar levels of fit to the best performing models in the field. Importantly for the purposes of this article, it was also demonstrated that the technique could be adopted to transform the feature values of a different distributional model type, namely the sparse implementation of the BEAGLE model of semantics (Recchia, Sahlgren, Kanerva, & Jones, 2015). Thus, the adoption of these transformations can be applied to features not directly based in word co-occurrence.

The modeling work described here will contrast the WW model with two different socially and communicatively oriented measures of language usage. All three models will be large matrices, transformed with the matrix transformation techniques proposed by Johns et al. (2019). Using the same modeling framework with three different models will allow for a determination of the different informational content that the three models are capturing. The two techniques developed by Johns et al. (2019) were entitled the global negative (GN) and distribution of associations (DOA) transformations. Johns et al. (2019) contains a complete description of these techniques, so the transformations will just be sketched here.

Although there are more sophisticated distributional models that have been used in cognitive modeling, such as word2vec, the distributional modeling framework of Johns et al. (2019) is advantageous for the purposes of this article. In particular, the goal of this article is to examine different types of information that can be used to train a distributional model, namely word-based co-occurrences and socially-based information. The transformations that will be described can be applied to any count-based occurrence matrix. Thus, the same model can be used to examine multiple models that are filled with different types of distributional information, while having identical processing mechanisms and cognitive assumptions. Following similar work (e.g., Levy, Goldberg, & Dagan, 2015), Johns et al. (2019) demonstrated that when the WW model was optimized with the various training methodologies that word2vec employs, a similar level of fit was found for the WW model compared to the best fitting word2vec models, putting the modeling framework used here on a solid footing in terms of the explanatory power of the framework compared to other distributional models. Some of the training methodologies will not be used here (e.g., the use of frequency subsampling), in an effort to keep the three proposed models equivalent in terms of model complexity, which may reduce the overall fit of the models. However, the goal of this article is not to generate the overall best-fitting model, but instead to test the contribution of different types of distributional information.

2.1. GN transformation

This technique adds an equal amount of positive and negative information into a word's representation. The first step of this transformation is to derive a vector of base-rate of occurrence across the columns (features) of the matrix, encoded in a global negative (GN) vector, by calculating the sum of each column in the matrix:

$$\mathbf{GN}_j = \sum_{i=1}^{n} \mathbf{M}_{i,j} \tag{1}$$

Where **GN** is the global negative vector, **M** is the word-by-word matrix, j is the column being calculated, and i increments through all n rows in the matrix. The values contained in the GN vector are related to the frequency of a word and the window size used by a model. The next step is to unit normalize the vector, so that the sum of the vector is 1, by dividing each entry in the GN vector by the vector's overall magnitude:

$$\mathbf{GN}_j = \frac{\mathbf{GN}_j}{\sum_{k=1}^{n} \mathbf{GN}_k}, \tag{2}$$

where k increments through each index in the **GN** vector.

The resulting normalized GN vector allows for an equal amount of negative information to be added into a word's representation, by taking the sum of a word's row, which represents the amount of positive information a word received during training and adding an equivalent amount of negative information. This is accomplished with the following equation:

$$\mathbf{M}_i = \mathbf{M}_i - (\mathbf{GN}^* \sum_{j=1}^{n} \mathbf{M}_{i,j}), \tag{3}$$

where $\mathbf{M}_i$ is a word's row in the matrix, and j goes through each column in the matrix. The end result of this transformation is that a word's representation contains an equivalent amount of positive and global negative information. If a value is positive in the resulting matrix, then it signals that the relationship between two words exceeds a base-rate of occurrence. If it is negative, then it signals that the word's co-occurrence patterns do not exceed a base-rate.

2.2. DOA transformation

This technique recognizes that the uniqueness of the connection between two words is already contained within the matrix. In order to capitalize on this realization, the columns in the WW matrix are transformed using z-scores:

$$\mathbf{M}_{i,j} = \frac{\mathbf{M}_{i,j} - \mu_j}{\sigma_j}, \tag{4}$$

where i represents a row in the matrix, j represents a column, μ_j represents the mean of the column, and σ_j represents the standard deviation of the column. The z-score indicates how many standard deviations the co-occurrence of two words is, over and above the other co-occurrence values contained in the columns of the matrix, providing a relative weighting for the association between two words.

However, the resulting scores are biased by word frequency, where words higher in frequency will have higher average association scores than lower frequency words. To normalize this, each value in a row is normalized with a z-score:

$$\mathbf{M}_{i,j} = \frac{\mathbf{M}_{i,j} - \mu_i}{\sigma_i}, \tag{5}$$

whereμ_i is the mean of a word's row, and σ_i is the standard deviation of that row. Thus, the DOA transformation happens in two steps – in the first step the columns are transformed to normal scores, and in the second step the rows are transformed into normal scores (in order reduce the impact of word frequency on a word's semantic representation). The result of this transformation are normalized association values, where the unique associations between words are highlighted.

2.3. Combined transformation

In Johns et al. (2019) it was found that the DOA transformation provided a superior fit to the GN transformation. However, the best fit was found when the GN transformation was applied first, followed by the DOA transformation, suggesting that they provide relatively unique changes to semantic space. In this article, initial analyses will contrast the GN, DOA, and combined transformations to determine if similar analyses hold for new model and corpus types.

2.4. Socially-based representations

The WW model as previously used by Johns et al. (2019) will be the model upon which the socially-based models will be compared to, as its performance is similar to other distributional models in the literature. The socially-based models will still be represented with a large matrix, but the features will not be words. Instead, they will consist of information about how users on Reddit communicate across discourses.

Following the work of Johns (2021), there will be two types of features used: users and discourse. Thus, the feature space will be different from the WW model. In particular, there will two types of models: 1) user-by-discourse (UD) and 2) discourse-by-user (DU). In the UD model, each feature is a user, thus the resulting memory matrix has a dimensionality of W $\times$ U, where U is the number of users contained in the Reddit data. In the DU model, each feature will be a discourse (subreddit), with the memory matrix having a dimensionality of W $\times$ D, where D is the number of subreddits contained in the Reddit data.

In order to make the social models unique from the WW model, the entries in the resulting matrix do not encode word frequencies. Instead, the entries will be communication patterns within the Reddit data. Both socially-based models encode how words are communicated across people and discourses. For the UD model, entries in the matrix will be the number of discourses that a user produced a word in. For the DU model, entries in the matrix will be the number of users who produced a word in a discourse. The use of a count of the number of discourses that a word was used in by a language user (the UD model), or the number of users who produced a word in a discourse (the DU model), is similar to the contextual diversity proposals of Johns (2021) and was chosen so that the features of these models are relatively dissimilar to the WW model.

The difference in the two models lie in how linguistic information is encapsulated. In the UD model, the strength of a feature is defined by the number of discourses that an individual user produced a word in, signaling how important that word is to a user for communicating across many different situations. In the DU model, the strength of a feature is defined by the number of users who used a word within a discourse, signaling how important that word is to communicating with others in that specific discourse. Subsequent analyses will determine their overlap and the relative importance of these feature types to accounting for word similarity.

In order to demonstrate how the representation for the UD and DU models are constructed, Fig. 1 contains a demonstration of how the two models are derived for three users (*Jennifer*, *Owen*, and *Lily*) and discourses (r/Nature, r/Sports, and r/Cooking) for the word *cardinals*. This figure shows that the two models have different featural constructions. For the UD model, the word representation is a count of the number of discourses a specific individual used a word in. This also entails that repeated usage of a word in the same discourse does not increase that word's feature strength (see Lily's feature value for an example). For the DU model, each user who produced a word in a given discourse increases that word's strength in that discourse column. When scaled to hundreds of thousands of users and tens of thousands of discourses, the resulting representation in these models will contain a significant amount of information about the communication patterns of different words. The above transcribed matrix transformation techniques will be applied to the WW, UD, and DU models, in order to improve the quality of their underlying representations.

Given that the WW model is measuring the co-occurrence rate of two words, it is expected that this model should provide the best performance to word similarity data, given the overlap between what the model is measuring and task requirements. However, the goal of this article is to determine if the socially-based UD and DU models account for any unique variance within the different datasets. If there is unique variance accounted for by these models, it would signal that how words are used during communication present unique information source to learn the meaning of words.

2.5. Reddit data and corpora

The corpora used in the following analysis was assembled from Reddit data made available from a website entitled pushshift.io (Baumgartner et al., 2020), which uses the publicly available Reddit API[1] to assemble the comments made on the site for each month and makes them available as database files. In this analysis, all comments made on the site from January 2006 to September 2019 were downloaded and extracted. From this data, two corpora types were constructed: user and discourse corpora. Each user corpus was composed of an individual user who produced at least 3000 comments. This criterion resulted in 334,345 user corpora. The discourse corpora contained each subreddit where these selected individuals communicated, resulting in 30,327 discourse corpora. The total number of words contained in both corpora sets is equivalent, with approximately 55 billion words being contained in each, but the information contained in them is organized differently (that is, by user or discourse). Each comment made was marked with either the discourse where a comment was produced (for the user corpora), or the user who produced a comment (for the discourse corpora). This extra-linguistic tagging enabled the training of the UD and DU models. More details on the information contained in these corpora can be found in Johns (2021).

2.6. Model training

The word list used to train the model was the 50,000 highest frequency words from the corpora. This means the WW model had a dimensionality of 50,000 $\times$ 50,00. For training the WW model, each sentence across the comments was used. Although it has been found that a small window size typically results in superior performance levels (e.g., Levy, Goldberg, & Dagan, 2015), here no window was set in order to not give the WW model an additional parameter (as there is no window parameter for the UD and DU models). Instead, word co-occurrences will be computed as to whether they occurred in the same sentence or not. All sentences in the corpora were used to train the WW model (the user and discourse corpora are identical in terms of sentences contained, so the WW model was trained on the user corpora).

For the UD model, the number of features (users) was capped at 80,000, in order to reduce computational requirements, as no performance increase was seen by adding additional users after this point. No mechanism of selecting users (e.g., only including the 80,000 most frequent commenters) exceeded the performance of randomly selecting users, so a random selection of users was used in the resulting model. Thus, the UD model has a dimensionality of 50,000 $\times$ 80,000. All subreddits were used as features, and so the DU had a dimensionality of 50,000 $\times$ 30,327.

Aside from negative sampling, another technological improvement introduced by neural embedding models is the use of word subsampling to mitigate the contribution of high frequency words during training. Subsampling works by using pre-processed word frequency values from the training corpus to incrementally skip high frequency words. Johns et al. (2019) demonstrated that subsampling increases the fit of the WW substantially, consistent with other analyses (e.g., Levy et al., 2015). However, given that the UD and DU models do not use word frequency values, it would be difficult to equate the models if a subsampling routine was used for the WW model. Instead, the stoplist from Landauer and Dumais (1997) was used to remove high frequency words for all three models. This stoplist consists of 441 high frequency words and consists of mostly function words. Any word pairs from the word similarity data that contained a word on the stop list were removed from the analysis.

2.7. Discussion

This section described the distributional modeling framework that is going to be used to explore the difference between two distributional model types: a classical word-based representation (the WW model) and two socially-based representations (the UD and DU models). The matrix transformations described in this section are useful as they can be applied to any feature-based representation

[1] Information on the API can be found at: https://www.reddit.com/dev/api/

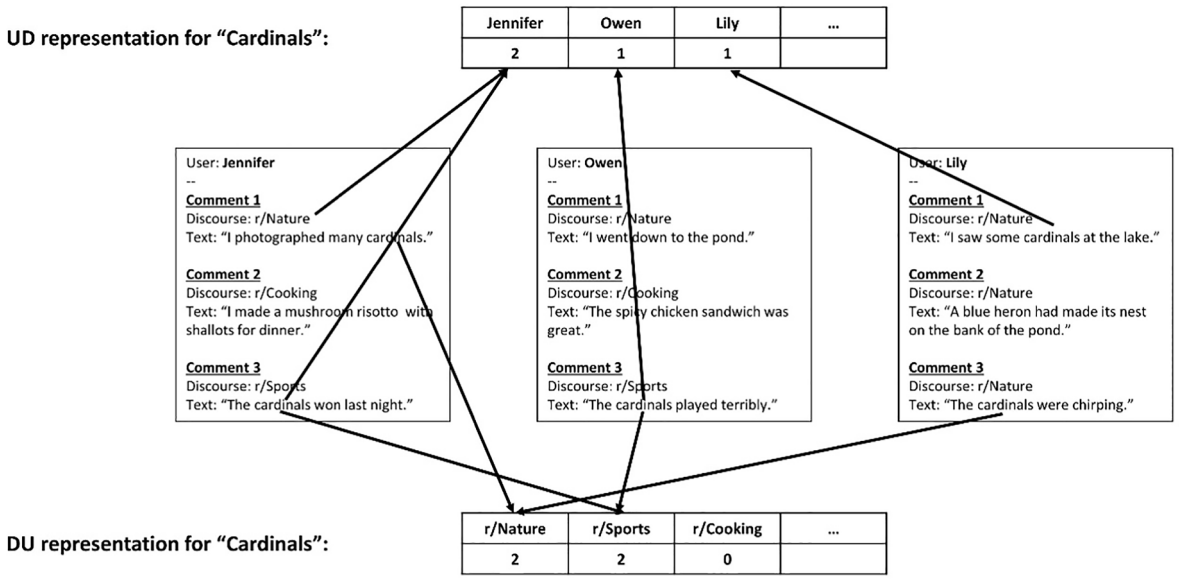

Fig. 1. An example of how the UD and DU models are constructed using examples from three hypothetical users and discourses. The UD model counts how many discourses an individual used a word in. The DU model counts the number of users who used a word in a specific discourse.

(as Johns et al., 2019 demonstrated). Thus, the different features proposed here – word co-occurrence frequency for the WW model, discourse communication pattern for the DU model, user communication pattern for the UD model – can be embedded within the same formal framework, and so can be directly compared to each other. The goal of the following analyses is not to posit that the UD and DU models are better accounts of lexical semantics than past approaches, but instead to determine if communication patterns of word usages provides a unique information source about word similarity over-and-above standard word co-occurrence learning models.

3. Results

3.1. Word similarity

In all models, similarity between words will be taken with a vector cosine between the two word's row in a memory matrix. The behavioral data that will be used to evaluate the three proposed models will be word similarity data, a standard task within distributional modeling and computational linguistics. In this task, a participant is given two words and is asked to rate how similar they feel those two words are on a given scale. Five different behavioral datasets are used in the analysis: (a) the WordSim data (n = 353; Finkelstein, et al., 2001), (b) RG1965 data (n = 65; Rubenstein & Goodenough, 1965), (c) the MTURK-771 data (n = 771; Halawi, Dror, Gabrilovich, & Koren, 2012), (d) the MEN data (n = 3000; Bruni, Boleda, Baroni, & Tran, 2012), and (e) the Radinsky-2011 dataset (n = 287; Radinsky, Agichtein, Gabrilovitch, & Markovitch, 2011). Table 1 contains a summary of these data for reference. Each dataset contains unique word pairs and, in some cases, different instructions for participants.

As a first pass at understanding the behavior of the three proposed models, Table 2 contains the Pearson correlation coefficients of the three models to each other for the different word sets from the word similarity datasets. This table shows that the models do produce cosine similarity values that are quite correlated with each other, as one would expect, but they are not perfectly correlated suggesting that they are capturing relatively different information about word meanings. The WW model tends to be most similar to the DU model, while the most similar two models are the UD and DU model, unsurprising given they are different organizations of the

Table 1
Description of the behavioral datasets utilized.

Type	Dataset	Label	N
Word similarity	Finkelstein et al. (2001)	WordSim	353
	Rubenstein and Goodenough (1965)	RG1965	65
	Halawi et al. (2012)	MTURK-771	771
	Bruni et al. (2012)	MEN	3000
	Radinsky et al. (2011)	Radinsky-2011	287
Lexical organization	Balota et al. (2007)	ELP_LDT, ELP_NT	40,467
	Keuleers et al. (2012)	BLDT	28,457
	Mandera et al. (2020)	OLDT	60,964
	Goh et al. (2020)	ALDT	10.155

Table 2
Similarity of three proposed models to each other using the various word sets.

Data	r(WW, DU)	r(WW, UD)	r(DU, UD)
WordSim	0.786	0.638	0.847
MTURK	0.781	0.611	0.849
RG1965	0.653	0.642	0.852
MEN	0.787	0.605	0.795
Radinsky	0.821	0.758	0.894
Average	0.766	0.651	0.847

Note. All correlation significant at the $p < .001$ level.

same data. Subsequent analyses will attempt to determine if the UD and DU models capture unique information about word similarity when directly compared to the WW model.

However, the word similarity datasets are handpicked word pairs chosen to contain a range of different levels of word similarity. By only assessing model similarity on these word pairs, the correlations between the models may be inflated. To provide an alternative assessment of the similarity of the three models, one million random word pairs were generated. This was done by taking the 5000 most frequent words (that did not occur on the stoplist) and then randomly selecting two words to form a single word pair, repeated one million times. Then the cosine similarity was taken between each word pair for all three models, and then correlations between the similarities across the three models was assessed. The correlation between the UD and DU model was found to be $r = 0.522$, $p < 0.001$, while the correlation between the UD and WW model was $r = 0.218$, $p < 0.001$, and the correlation between the DU and WW model was $r = 0.238$, $p < 0.001$. This suggests that while the DU and UD are moderately correlated to each other, they do offer quite distinct similarity patterns, when random semantic similarity is considered, and both are relatively distinct from the WW model.

As a first demonstration that DU and UD models are capturing semantic information, the Test of English as a Foreign Language (TOEFL), originally used by Landauer and Dumais (1997), was used. The TOEFL is a synonym test that consists of a target word and a set of 4 alternatives where one word is a synonym of the target, and the goal of the task is to find the synonym (e.g., for the target word *zenith*, choose *pinnacle* when presented in a list with *completion*, *outset*, and *decline*). There are 80 questions, and the test is scored by determining the number that the model got correct. Chance in this test is 25% as there are four possible alternatives for each question. This is a more difficult test of the DU and UD model because many of the words used are low in frequency and quite abstract, and these models do not directly encode word-word relations. However, this test will provide a first test as to whether the model can account for semantic behavior. The combined transformation was used for all three models. The WW model got 65 questions correct (81.25%), replicating the strong performance found in Johns, Mewhort, and Jones (2019). The DU model got 51 questions correct (63.75%) while the UD model got 39 questions correct (48.75%). Both of the newly proposed models are above chance (which is 25%), suggesting they are capturing semantic information, but have worse performance than the WW model. The following analyses using word similarity data will elucidate the performance of the different models and determine if the socially-based representation types account for unique variance in these data.

In evaluating models of word similarity, it is common to take the Spearman correlation between the data and the model's similarity values. A Spearman correlation transforms both the data and the model's output into ranks before assessing the correlation. This type

Table 3
Pearson correlation coefficients of the proposed distributional models to the five word similarity datasets.

		Transformation			
Model	Data	None	GN	DOA	Comb
WW	WordSim	.057 *n.s.*	0.408 ***	0.542 ***	0.58 ***
	MTURK	.068 *n.s.*	0.237 ***	0.517 ***	0.547 ***
	RG1965	−.07 *n.s.*	.242 *n.s.*	0.523 ***	0.55 ***
	MEN	0.045 **	0.378 ***	0.635 ***	0.669 ***
	Radinsky	.042 *n.s.*	0.391 ***	0.558 ***	0.592 ***
	Average	0.028	0.331	0.555	0.588
DU	WordSim	.067 *n.s*	0.217 ***	0.442 ***	0.585 ***
	MTURK	0.102 **	0.203 ***	0.488 ***	0.535 ***
	RG1965	.174 *n.s.*	0.336 **	0.475 ***	0.581 ***
	MEN	0.14 ***	0.229 ***	0.582 ***	0.611 ***
	Radinsky	.008 *n.s.*	0.119 *	0.566 ***	0.609 ***
	Average	0.098	0.221	0.511	0.584
UD	WordSim	0.12 *	0.421 ***	0.342 ***	0.438 ***
	MTURK	0.129 ***	0.389 ***	0.418 ***	0.402 ***
	RG1965	.212 *n.s.*	0.386 ***	0.392 **	0.49 ***
	MEN	0.138 ***	0.403 ***	0.461 ***	0.378 ***
	Radinsky	0.344 ***	0.502 ***	0.547 ***	0.576 ***
	Average	0.189	0.42	0.432	0.457

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$; *n.s.* = not significant.

of correlation is used in distributional modeling as it typically results in a better fit. However, it also removes natural variation from the data and model performance as word similarity is typically not linear (Johns & Jones, 2010), but a rank transformation forces a linear organization upon the data. In order to determine what the impact of a rank transformation has upon the fit of the data, both a Pearson and Spearman correlation will be used to assess model performance in the coming analysis. Table 3 contains the Pearson correlation coefficients of the various models, across the three different matrix transformations (GN, DOA, and combined) to the different word similarity datasets, while Table 4 contains this same data but utilizing a Spearman correlation.

Importantly, these tables show that both of the socially-based models achieve significant levels of correlations using the combined transformation, with the DU model achieving a similar average level fit as the WW model when a Pearson correlation is used. This suggests that using the socially-based features, word meanings can still be captured.

These tables also show that the advantages of the GN and DOA transformations of the WW model replicate the findings of Johns et al. (2019), where the GN and DOA transformation both provide major improvements over the untransformed matrix. Also replicating the results of Johns et al. (2019), the DOA transformation is the superior transformation type, while the two combined transformations offer the best fit. This pattern of fit mostly extended to the UD and DU models, with some variations. For the DU model, the DOA transformation performed relatively better than the GN transformation, compared to the WW model, using both a Pearson and Spearman correlation. The UD model showed an opposite pattern when using a Pearson correlation where the GN transformation performs better than the DOA transformation, and the two transformations had roughly similar performance levels when using a Spearman correlation. These differences between representation types and transformations suggests that the distributional properties of a representation impact the utility of the different matrix transformations. However, for all three models, across the two correlation types, the combined transformation offered the best performance level, consistent with the results of Johns et al. (2019).

Tables 2 and 3 show that the WW model sees a much larger average improvement in performance using the Spearman correlation compared to the Pearson correlation (from an average correlation of 0.588 to 0.721), compared to the DU model (from 0.584 to 0.618) and UD model (from 0.457 to 0.507). With the Pearson correlation, the WW and DU model offer an equivalent fit. This provides evidence that ranking the similarity values and data provides a benefit to the WW model, compared to the other two model types. Given that a Pearson correlation is the default correlation type for continuous variables (which both word similarity data and cosine similarity from distributional models are), non-ranked data will be used in subsequent analyses. This is due to the use of rank transformations providing an advantage to the WW model, as well as limiting the amount of variance in the data. The issue of variance in data is especially true for word similarity data, given the limited size of the word similarity datasets (especially compared to the size of the datasets used in lexical organization), which may limit the ability of a model to account for unique variance. Subsequent analyses will focus on separating the contributions of the different models using regression analyses. To keep all models equivalent, the combined matrix transformation will be used for each in the coming analyses.

In order determine if the UD and DU account for any unique variance in the word similarity data, hierarchical linear regression analyses will be employed. This type of analysis has been repeatedly employed in studies of lexical decision and naming data (e.g., Adelman, et al., 2006; Johns, et al., 2016) The end result of this analysis technique is the amount of predictive gain (measured as percent ΔR^2 improvement) for one predictor over other competing predictors, when they are contained in a regression.

The first regression analysis will calculate the amount of unique variance that each variable accounts for in the different word similarity datasets, when combined in a regression. The results of this regression are contained in Table 5, which shows that both the WW and DU models account for significant levels of variance across all datasets, while the UD model accounts for significant variance

Table 4
Spearman correlation coefficients of the proposed distributional models to the five word similarity datasets.

		Transformation			
Model	Data	None	GN	DOA	Comb
WW	WordSim	0.402 ***	0.407 ***	0.711 ***	0.73 ***
	MTURK	0.185 ***	0.23 ***	0.649 ***	0.658 ***
	RG1965	.201 *n.s.*	0.277 *	0.721 ***	0.765 ***
	MEN	0.293 ***	0.376 ***	0.759 ***	0.767 ***
	Radinsky	0.306 ***	0.397 ***	0.699 ***	0.689 ***
	Average	0.277	0.337	0.707	0.721
DU	WordSim	.099 *n.s.*	0.225 ***	0.509 ***	0.645 ***
	MTURK	0.131 **	0.211 ***	0.51 ***	0.557 ***
	RG1965	0.263 *	0.382 **	0.532 ***	0.614 ***
	MEN	0.186 ***	0.241 ***	0.628 ***	0.662 ***
	Radinsky	.075 *n.s.*	.11 *n.s.*	0.554 ***	0.614 ***
	Average	0.151	0.234	0.546	0.618
UD	WordSim	0.137 *	0.451 ***	0.396 ***	0.488 ***
	MTURK	0.132 **	0.402 ***	0.428 ***	0.435 ***
	RG1965	.185 *n.s.*	0.426 **	0.423 **	0.561 ***
	MEN	0.152 ***	0.441 ***	0.487 ***	0.458 ***
	Radinsky	0.36 ***	0.468 ***	0.556 ***	0.594 ***
	Average	0.193	0.437	0.458	0.507

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$; *n.s.* = not significant.

in three of the datasets (the three with the largest sample sizes, which may suggest that there may not be enough data contained in the other datasets to find a significant signature of this model). The significance values in this table are F-tests evaluating whether the addition of a variable provided a significant increase in variance accounted for over the other nested variables. Both the WW and DU model accounted for considerable amounts of unique variance, with the WW model accounting for slightly more variance on average than the DU model. This suggests that the DU model is capturing a unique component of word similarity. There is considerable variability in the amount of variance that the different variables account for, likely due to differences in sample size of the different datasets, as well as the different word pairs that are used.

The second regression will exclude the UD model in the calculation of unique variance, given that it accounted for the least amount of unique variance of the three proposed models. Additionally, as Table 2 shows, the UD and DU models are quite correlated with each other for these datasets, and so the inclusion of the UD model in the results contained in Table 5 may underestimate the contributions of the DU model. The results of this regression are contained in Table 6. This table shows that the exclusion of the UD model did not cause a major change in variance accounted for. Overall, the WW still accounted for slightly more average variance compared to the DU model (11.83% to 11.07%), while the DU model accounted for the most variance in three of the five datasets (although the three smallest datasets). Thus, both the WW and DU account for significant variance in the word similarity dataset, signalling that these offer complementary sources of information about the meaning of words. This suggests that the communicative patterns that the DU model encodes provides a unique measure of word similarity, which standard distributional models that learn only from word-word co-occurrence patterns do not encode in their representations.

To further explore differences in the representations of the WW and DU model, Table 7 contains the ranked nearest neighbors to four words (*poodle*, *basketball*, *chemistry*, and *bird*) for both model types. This table shows that there is a different pattern of semantic neighborhoods for these words, with the WW representation tending to be similar to other/subordinate semantic category members (e.g., different dog breeds for *poodle*, different sports for *basketball*, different scientific fields for *chemistry*, and different types of bird species for *bird*). In contrast, the DU model tends to create semantic neighborhoods that are filled with properties of the words (e.g., *grooming* and *matting* for *poodle*, *rebounds* and *athleticism* for *basketball*, *undergrad* and *grades* for *chemistry*, and *beak* and *feathers* for *bird*). This suggests a qualitative difference in the type of knowledge that these two representations are learning, with the WW model extracting item-to-item information and the DU model extracting item-to-discourse level information.

To test this hypothesis, a simulation was run using a simplified semantic categorization task with 35 categories extracted from the van Overschelde, Rawson, and Dunlosky (2004) category norms. In this simulation, both the WW and DU model will be given a category member (e.g., *cucumber*) and the models will have to assign a category label for that member, equivalent to a 35-AFC task. The models will base this decision on two different similarity comparisons: 1) to the category labels (e.g., *vegetables*) or 2) the most prototypical category member (e.g., *carrot*). The most prototypical category member will be the most frequently produced member of each category, and this word will not be categorized. This resulted in 690 words from the 35 categories that will be used as task stimuli. Based on the patterns found in Table 7, it would be expected that the DU model should perform better when given category labels as compared to the prototypical category member, while the WW model should show less of a difference. For each model type, the performance of the model will be assessed across all 690 words and analyzed at the item-level.

The results of this simulation are contained in Fig. 2 and demonstrate that both models achieved quite high performance (chance in this task is 1/35 or approximately 2.85%; the DU model achieved a performance of 69.28% while the WW model was at 68.7% using the category label similarity). This figure shows that for the WW model there is no statistical difference in the performance of the model when using either the category label or prototypical category member to derive similarity [$t(689) = 0.567$, n.s.]. However, there was a large difference found for the performance of the DU model, such that the model performed much better when given a category label than a prototypical category member [$t(689) = 8.27$, $p < .001$]. For both representation types, the models performed best when given a category label, but there was no statistically significant difference in model performance for the WW or DU model [$t(689) = 0.34$, *n.s.*]. However, across all of the words tested, model performance was only moderately correlated, with an $r(690) = 0.533$, $p < .001$, suggesting that there is considerable variance in the words that the two models are able to successfully categorize.

Although this simulation is rather ad-hoc, it does serve to demonstrate that there are substantial underlying differences in the types of information that the two representation types are deriving from their training materials. This validates the previous analyses on word similarity data demonstrating unique variance accounted for by the two model types. The next section will determine if these unique contributions hold for the two models when fitting to lexical organization data, through the examination of the distribution of feature strength for the different proposed models.

3.2. Feature strength

Lexical representations have two properties – phase and magnitude (Jones, Dye, & Johns, 2017). Phase represents the meaning of a word, and how in-phase two words are is typically measured with a similarity value derived with a vector cosine. The previous analyses established that the socially-based representations (especially the DU model) contains unique information about the phase, or meaning, of two words compared to standard distributional models as embedded in the information encoded by the WW model. Magnitude represents the strength of a word in the lexicon. In terms of distributional modeling, the Semantic Distinctiveness Model (SDM) of Jones et al. (2012) and Johns (2021) has explored the contribution of distributional statistics on lexical strength and has established that distributional properties of word occurrence matter to the organization of the lexicon.

An underexplored aspect of distributional models and lexical organization is feature strength, or the impact that having well defined features in a word's semantic representation has on its strength in the lexicon. This is due to most distributional models having representations where the signature of word meanings is distributed across a word's representation (e.g., across a hidden layer for

Table 5
Amount of unique variance each model accounts for in the different word similarity datasets.

	ΔR^2 in %		
Data	WW	DU	UD
WordSim	8.86***	13.92***	4.05**
MTURK	1.06***	7.62***	1.53*
RG1965	13.19*	13.21*	.51 *n.s.*
MEN	14.8***	1.2***	7.6***
Radinsky	6.28***	2.26*	.5 *n.s.*
Average	8.838	9.442	2.838

Note. * = p < .05; *** = p < .001; *n.s.* = not significant.

Table 6
Amount of unique variance the WW and DU models account for in the word similarity datasets.

	ΔR^2 in %	
Data	WW	DU
WordSim	9.49***	11.34***
MTURK	11.45***	7.43***
RG1965	12.66*	21.71**
MEN	19.26***	3.25***
Radinsky	6.31***	11.62***
Average	11.83	11.07

Note. * = p < .05; *** = p < .001.

Table 7
Ranked nearest neighbors across four words for the WW and DU models.

Poodle		Basketball		Chemistry		Bird	
WW	DU	WW	DU	WW	DU	WW	DU
Shih	Groomer	Soccer	Rebounds	Biochemistry	Biochemistry	Hoatzin	Beak
Maltese	Matting	Netball	Coach	Physics	Biology	Lorikeet	Feathers
Schnauzer	Puppy	Football	Athleticism	Neuroscience	Undergrad	Bowerbird	Parrots
Terrier	Purebred	Sport	Fouls	Electrochemistry	Grades	Jacana	Plumage
Breeds	Breeds	Hockey	Rebounder	Science	Sciences	Caracara	Finches
Hypoallergenic	Housebreaking	Baseball	Threes	Paleontology	Lab	Jacamar	Avian
Paddled	Matted	Korfball	Sports	Geology	Graduate	Oilbird	Beaks
Retriever	Puppies	Team	Coaches	Biology	Undergraduate	Trogon	Starlings
Wheaten	Breeder	Handball	Championship	Inorganic	Molecular	Ovenbird	Nestlings
Miniature	Pup	Softball	Coaching	Anthropology	Career	Oystercatcher	Songbirds
Spaniel	Litters	Tennis	Championships	Radiochemistry	Chemists	Lapwing	Fledglings
Purebred	Topknot	Athletic	Playoffs	Engineering	Chromatography	Parrot	Cockatiel
Terriers	Retriever	Racquetball	Refs	Stereoisomerism	Molecule	Myna	Lovebird
Chihuahua	Breeders	Volleyball	Football	Lab	Professors	Whitethroat	Perches
Spaniels	Barking	Lacrosse	Playoff	Epidemiology	Chemical	Lovebird	Macaw

neural embedding models), with individual feature magnitudes not being overly informative about the strength of a word in memory. However, one of the advantages of the Johns, Jones, and Mewhort (2019) distributional modeling framework is that it maps all feature values into z-scores that represent the strength of each feature for a word, relative to that feature's strength for other words in the lexicon. The magnitude of a word's feature is thus interpretable, with a larger feature value signaling that the feature is more informative or important for the meaning of that word. This entails that the three specified models (WW, UD, and DU) can be directly compared to each other in terms of the feature strengths that each word has and can be used to determine if differential feature strengths account for unique variance in the retrieval time of words from the lexicon, a unique analysis in distributional semantic modeling. It is assumed that words with stronger semantic features will be easier to retrieve from memory, as they would have more distinctive features compared to words that have more neutral feature strengths. This is similar in principle to the proposals of the Retrieving Effectively from Memory (REM) model of Shiffrin and Steyvers (1997).

To demonstrate what the feature strength distribution for these models look like, Fig. 3 contains a histogram of the feature strength models for the DU model. This figure only contains the proportions of occurrence for z-scores from −0.4 to 0.4, as the vast majority of feature values fell between these criteria (approximately 92%). This figure shows that most feature strengths are below zero (approximately 82.3%), which signals that the positive semantic features of a word provide most of the distinctive features for the meaning of a word. To examine the impact of strong feature values, a criterion was set in order to calculate the number of features that

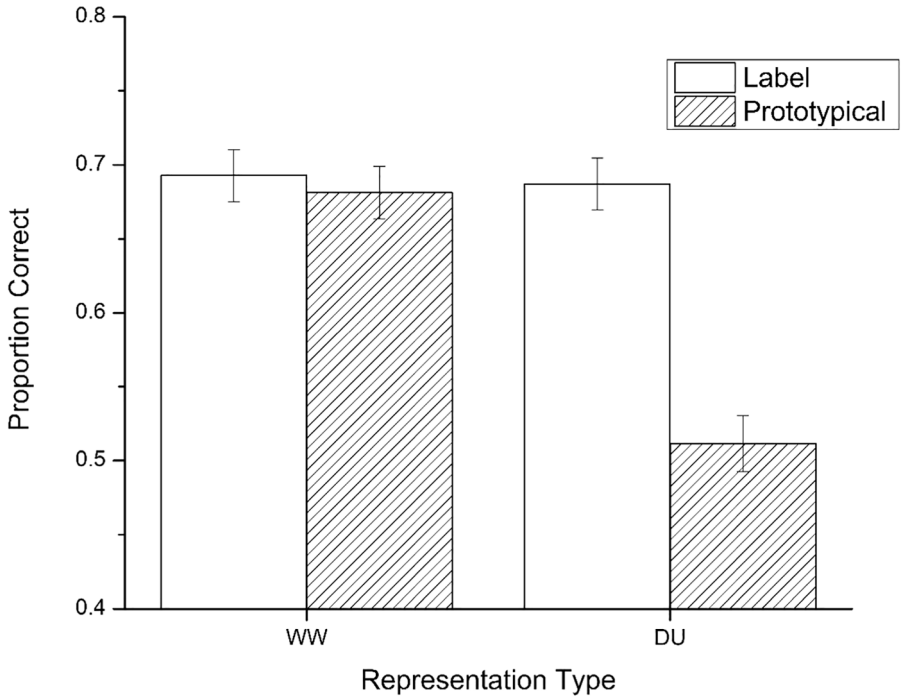

Fig. 2. The results of the semantic categorization simulation using both the WW and WU representations and using either the category label or prototypical category member similarity. This simulation demonstrates that the two representations are extracting different types of information from the lexical environment leading to different model performance.

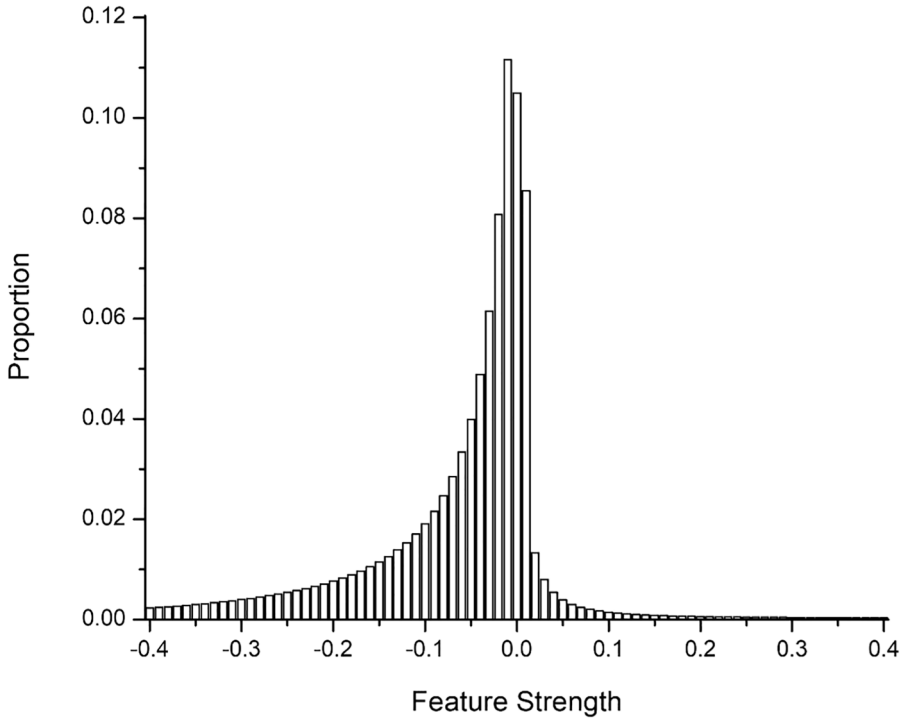

Fig. 3. A histogram of feature strengths for the DU model.

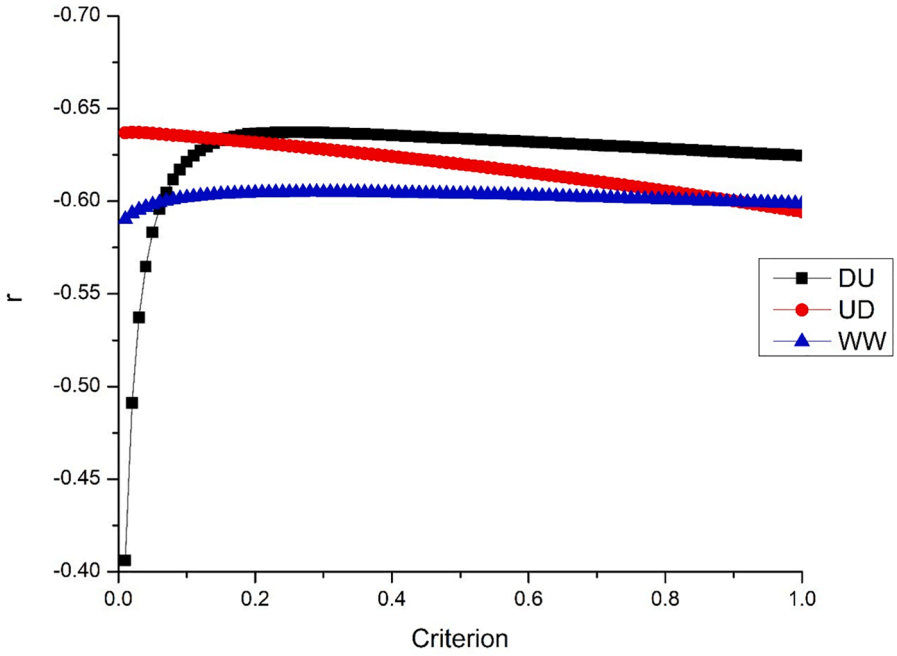

Fig. 4. Impact of the feature strength criterion on the fit of three models to lexical decision reaction time data from the English Lexicon Project (Balota et al., 2007).

exceeded that criterion for each model. The resulting variable will be entitled the feature strength of a model and will be a count of the number of features that exceeds the criterion. The resulting feature strength count was reduced with a logarithm, as this was found to increase model performance across all three model types.

Instead of examining semantic similarity data as was done in the previous analyses, this section will examine mega-datasets of reaction time data for lexical decision and naming tasks to determine the impact of a word's feature strength on a word's storage in memory. Similar data have previously been shown to be influenced by similar aspects of distributional models, such as semantic neighborhood size (Shaoul & Westbury, 2010a). In particular, five reaction time datasets will be examined: 1) English Lexical Project decision and naming time data (ELP_LDT and ELP_NT; Balota, et al., 2007), 2) British Lexicon Project lexical decision time (BLDT; Keuleers, Lacey, Rastle, & Brysbaert, 2012), 4) the recently released response times an online lexical decision task (OLDT; Mandera, Keuleers, & Brysbaert, 2020), and 5) recently released auditory lexical decision time data (ALDT; Goh, Yap, & Chee, 2020). Table 1 contains a summary of these data for reference. All reaction time data was z-transformed. Additionally, the probit-transformed word prevalence data from Brysbaert, Mandera, McCormick, and Keuleers (2019) will be used to examine the impact of feature strength on the probability of a person knowing a word. Multiple datasets were utilized to ensure that any findings generalized across the different types of data. To account for this data, the model's vocabulary was increased to 80,815 words, in order for all of the data points in these datasets to be accounted for, with a stoplist not being used in this analysis.

To determine the impact of the feature strength criterion on the resulting fit of the three models to the different lexical organization data, a simulation was run calculating the correlation of the number of features to the lexical decision RT data from the ELP. To accomplish this, the feature strength criterion was manipulated for the three model types in steps of 0.01. The result of this simulation is contained in Fig. 4 and shows that there is a different pattern of results across the models. For the DU model, there is a consistent increase in model performance until plateauing, while for the WW and UD models maximal model performance is found with a criterion around zero. Only the DU model was overly influenced by the setting of the strength criterion. The reason for this is likely due to dimensionality of the DU model, which is considerably smaller (and thus has fewer features) than the other two models. The UD and WW also have more sparse features (meaning most entries in the model's matrices are set at zero), which Johns, et al. (2019) showed resulted in most normalized values being shifted to being negative features (as zero is below the mean feature value, so the resulting normalized value is a negative z-score). To validate this hypothesis, it was found that for the GN + DOA DU model, 17.4% of features were positive, while for the UD model with these transformations it was found that 8.82% of features were positive, and for the WW model 8.67% were positive. Thus, for the DU model to be as discriminative as the WW and UD model, it needs a higher strength criterion which results in this model having a higher optimal criterion value.

To ensure that model performance is maximized, an individual feature strength criterion was set for each model and each dataset

Table 8
Correlations between the feature strength values and various reaction time datasets.

Data	WW	DU	UD
ELP_LDT	−0.605	−0.606	−0.637
ELP_NT	−0.515	−0.524	−0.548
BLP_LDT	−0.613	−0.635	−0.635
WP_LDT	−0.719	−0.751	−0.703
ALDT	−0.483	−0.53	−0.538
WP	0.692	0.73	0.699
|Average r|	0.604	0.629	0.627

Note. All correlations significant at the $p < .001$ level.

Table 9
Amount of unique variance the feature strength that each model accounts for over the WF and UCD-SD-PR variables for the five lexical organization datasets.

	ΔR^2 in %		
Data	WW	DU	UD
ELP_LDT	2.409***	4.705***	1.62***
ELP_NT	2.67***	4.21***	1.35***
BLP_LDT	0.631***	1.87***	1.25***
WP_LDT	0.017***	0.813***	0.024***
ALDT	3.75***	4.871***	0.3**
WP	1.497***	1.986***	0.336***
Average	1.829	3.076	0.813

Note. ** = $p < .01$; *** = $p < .001$

independently based on a Pearson correlation[2]. Table 8 contains the resulting correlations between the optimal feature strength values for the three models and the five datasets. This table shows that all models have a strong correlation to the reaction time datasets. This finding confirms that words that have stronger semantic features in memory are faster to be retrieved from the lexicon. Overall, the DU and UD models exceed the performance of the WW model on average, suggesting that the socially-based models outperform the word-based distributional representation in accounting for lexical organization data, similar to the results of Johns (2021).

However, the feature strength values are confounded with other lexical strength measures, such as word frequency and contextual diversity measures. For example, the DU feature strength values have an $r = 0.923$, $p < .001$ to WF and $r = 0.945$, $p < .001$ to the best fitting CD measure from Johns (2021; the overall best fitting measure was the UCD-SD-PR metric). To determine whether the feature strength values account for unique variance over and above the WF and UCD-SD-PR metrics, a regression was done calculating the amount of improvement in variance accounted for when these two variables are controlled for. The result of this analysis is contained in Table 9, which shows that for all datasets, each of the models accounts for significant levels of unique variance over the WF and CD variables. Although the overall amount of unique variance that the models account for is less than that in the word similarity datasets, this is tempered by the fact that the lexical organization datasets are order of magnitudes larger than the word similarity datasets (and is consistent with model fits for CD measures over WF; Adelman et al., 2006; Adelman & Brown, 2008; Johns, Dye, & Jones, 2020). This finding suggests that the feature strength counts from the various models do account for unique variance in the lexical organization datasets, implying that the magnitude of the strength of features has psychological reality in the storage of words in the lexicon. Overall, the DU model, on average, accounted for the most unique variance, followed by the WW and UD models. Similar to the results of the previous analyses on semantic similarity ratings, the DU model offers the better accounting to the data compared to the UD model for the socially-based representations.

In order to determine if the DU model provides an overall better fit to the lexical organization data than the WW model, an equivalent regression to that contained in Table 6 was done, where the amount of unique variance that the DU and WW model accounts for over and above each other was calculated. However, in this analysis the WF and UCD-SD-PR variables were also controlled for. The result of this analysis is contained in Table 10 and shows that the feature strength values from the DU model account for substantially more unique variance than the WW model when compared directly to each other. This provides an additional source of data that the semantic representations derived from the DU account for unique variance in lexical data, in compliment to standard distributional models that are based only in word co-occurrence statistics.

4. General discussion

The goal of this article was to determine if social communication patterns of word usage provide unique insight into word

[2] The feature strength for the different models with feature strength criterion of 0.2 are available at: https://btjohns.com/distributional_feature_strength.xlsx

Table 10
Amount of unique variance the WW and DU models account for in the lexical organization datasets over each other, with the WF and UCD-SD-PR variables also controlled for.

	ΔR^2 in %	
Data	WW	DU
ELP_LDT	1.04***	2.604***
ELP_NT	0.207***	1.452***
BLP_LDT	0.126***	1.493***
WP_LDT	0.013*	0.813***
ALDT	1.633***	2.816***
WP	0.657***	1.156***
Average	0.612	1.723

Note. * = $p < .05$; *** = $p < .001$

meanings, using established tools from distributional semantic modeling. In order to accomplish this, the matrix transformation techniques described in Johns et al. (2019), which adopted distributional training techniques from neural embedding models (Mikolov et al., 2013), was used in order to assess the different roles that word-based co-occurrence and communication statistics contribute to lexical similarity and strength. The advantage of this model type is that different feature-based representations (from word co-occurrence values to different instantiations of user and discourse communication patterns) can be contrasted with each other, while using the same underlying modeling framework.

Three matrix models of lexical semantics were tested. The standard distributional approach was represented with a Word × Word (WW) matrix where each word's meaning was represented by its co-occurrence frequency with other words, similar to count-based models of lexical semantics (e.g., Bullinaria & Levy, 2007, 2012). The first socially-based representation type used a User-by-Discourse (UD) representation, where the features (columns) of the distributional matrix is a user from Reddit, and each element in the matrix is the number of different discourses that a user produced a word in. The second socially-based representation used a Discourse-by-User (DU) representation type, where the feature in the matrix is a discourse (subreddit), and each element in the matrix is the number of users who produced a word in that discourse.

It was found that both social representation types provide a significant correlation to word similarity data and performed above chance on a synonym test. When combined with the WW model in a regression analysis, it was found that the DU model accounted for significant levels of unique variance across all five different datasets used, while the UD accounted for significant levels of unique variance on three of the five datasets. A second regression directly comparing the WW and DU models found that both models account for significant amounts of unique variance across all datasets, with the WW accounting for slightly more unique variance on average, with the DU model accounting for the most unique variance in three of the five datasets. This finding suggests that the DU model is capturing unique information about word meanings outside of word-word co-occurrence statistics. A further simulation examining semantic categorization found that the two models performed similarly but did so using considerably different types of information.

To further evaluate whether these models account for unique variance in lexical processing data, the feature strength of the differing models was used to evaluate the fit of the models to lexical organization data. The ability to conduct this analysis is due to the modeling framework of Johns et al. (2019) generating semantic features with interpretable magnitudes. Overall, it was found that words that have a greater number of stronger features in memory are retrieved faster from the lexicon, with the socially-based feature strengths providing a better accounting than the word-based feature strengths, a unique finding in distributional modeling, and complementary to other findings examining the role of social information in lexical organization and memory processing (Johns, 2021, in press). Multiple regression analyses found that DU model accounted for the most unique variance across all datasets, over and above the other tested models and various lexical strength measures, with the WW model also accounting for significant levels of unique variance. This result provides converging evidence, together with the results of the word similarity simulations, that the semantic features derived from the DU provide unique information about a word's lexical representation outside of those captured by word co-occurrence statistics, which form the basis of standard models of distributional semantics.

The success of the DU model signifies that the best socially-based information source for deriving word meanings is based around the topic of conversation where language is being used, with the strength of the discourse features being determined by a count of the number of people who used a word in that discourse. Counting the number of users who produced a word in a discourse indicates the importance of that word to communicating effectively within that topic of conversation. If everyone uses a particular word to communicate within a discourse, then it is likely that you will also need to use that word. Thus, understanding the meaning of that word and how it is used across people and discourses is fundamental to communicative effectiveness.

The finding that communication patterns should be organized within discourse (in contrast to within user) is different from the results of Johns (2021) when examining lexical organization data, where it was found that the best fitting model consisted of a count of the number of users who produced a word, modified with the semantic distinctiveness model (Johns, 2021). This suggests that different types of lexical behavior use different sources of lexical information to varying degrees (as demonstrated in the simulation contained in Fig. 2). A goal in the computational modeling of language should be to develop integrated models of language, where multiple types of lexical processes are couched in the same general framework, with different weights given to different types of information. It is likely the case that different environmental information sources (e.g., user versus discourse word usage) are utilized to different extents across different tasks (e.g., lexical organization versus lexical semantics). This entails that combined models will

need to differentiate multiple types of information to account for different data types from the same model.

Classic distributional models of semantics capitalize on the redundancy of word co-occurrences across contexts and have been proved to be highly effective at inferring word meanings through this statistical source. However, language is not separate from other aspects of human experience. Language is an intentional behavior with communicative purpose, whether spoken or written. The results of this article suggest that the discourse where a word is produced, and who produced that word, offers unique insight into the meaning of a word. That is, each occurrence of a word is not equivalent – there is extra-linguistic information that surrounds the usage of a word, which should also be considered by distributional models. From a statistical learning point of view, this suggests that lexical statistics are not just tied to linguistic symbols, but also that surrounding contextual information is learned and used in language learning and processing.

To further evaluate the importance of social information in the acquisition of word meanings, empirical studies will be necessary. A promising avenue to evaluate the importance of social information at the beginning of word learning is provided by recent work mixing artificial and natural language learning (e.g., Johns, Dye, & Jones, 2016; Mak, Hsiao, & Nation, 2021; Snefjella, Lana, & Kuperman, 2020), where pseudowords are introduced into natural language contexts to examine how new word meanings are acquired. By introducing communicative factors into this type of paradigm (e.g., have agents produce the language rather than a passive reading task), it would be possible to manipulate social factors in language learning. The work described here could provide concrete predictions about performance in such a paradigm, for example if only one agent repeatedly uses a word the feature strength for that word should be less than if many people used a word only once. Such empirical work would further enable an understanding on the interaction of different sources of distributional information on word meaning acquisition.

The importance of extra-linguistic information has been established in grounded approaches to cognition (Barsalou, 1999, 2008), where the perceptual referents of language are intimately tied to language comprehension and usage. Indeed, initial criticisms of the distributional approach to lexical semantics (represented with the classic LSA model of Landauer & Dumais, 1997) focused on the lack of perceptual information in the model's representation (e.g., Glenberg & Robertson, 2000). However, no theory can account for all data, especially those at the beginning stages of development. The work of Landauer and Dumais (1997) established the promise of integrating advanced computational techniques into cognitive modeling, which should serve as a guidepost for future research into big data and machine learning work in the cognitive sciences (Johns et al., 2019). Subsequent models have demonstrated the ability of distributional models to integrate of perceptual information into the representations constructed with distributional models (Andrews, Vigliocco, & Vinson, 2009; Lazaridou et al., 2017; Bruni et al., 2014; De Deyne, Navarro, Collell, & Perfors, 2021; Johns & Jones, 2012; Riordan & Jones, 2011), demonstrating the flexibility of these model types in integrating extra-linguistic information sources into their representations.

Distributional models that integrate grounded information into their representation acknowledge the limitation of having an amodal representation of language (Andrews, Vigliocco, & Vinson, 2009; De Deyne et al., 2021; Riordan & Jones, 2011). However, focusing purely on words and the perceptual environment in which they occur ignores another information source: who produced that word and in what discourse. The results of this article point to the importance of social information in computing lexical semantic representations, following usage-based and adaptive theories of language processing (Beckner et al., 2009; Christiansen & Chater, 2008; Tomasello, 2003, 2009). These theories propose that language is fundamentally communicative, which entails that in order to construct more realistic models of language, distributional models should begin to take the social world into account in their computations.

Usage-based and adaptive theories of language propose that who produces an utterance, and in what context, is an important environmental cue that aids the general learnability of language. Scaled to adult levels of linguistic experience, these theories posit that communicative patterns of word usage should be integrated into the lexicon. Johns (2021) established that communicatively-oriented measures of lexical strength account for considerably greater levels of unique variance in lexical organization data compared to classic measures, such as word frequency, establishing the viability of integrating social information into computational models of language. The work described here extends these results to lexical semantics and suggests that communicative information is a general property of lexical statistics, which should be integrated into future models of language processing.

The goal of this work was not to invalidate the mechanisms of previous distributional models, but instead to broaden the application of this model type, in order to ground the models within the social world. The production and comprehension of language is embedded within multiple types of contextual information, from who produced that language, the source of that language (e.g., newspaper article, a lecture, twitter, facebook), how that language was experienced (i.e., spoken versus written), among other granular aspects of linguistic context. The majority of research within distributional modeling has focused on exploiting different aspects of word co-occurrence statistics, which has repeatedly been shown to be an extremely powerful source of information about word meanings. However, to grow these models to be more realistic models of language processing requires integrating other sources of information into their representations. The models explored here are capable of accomplishing this, but in an inelegant matter. Additionally, the vector representations derived using the methods in this article are difficult to combine into a single representation, as the vectors cannot be just concatenated together, because of issues of dimensionality and the differences in distributional statistics across the different representation types. Future modeling efforts should focus on how to construct this integration in a more parsimonious fashion.

Some distributional modeling frameworks may be easier to integrate social information into than others. For example, corpus-based instance theories of language (Crump, Jamieson, Johns, & Jones, 2020; Kwantes, 2005; Jamieson, Avery, Johns, & Jones, 2018; Johns & Jones, 2015; Johns, Jamieson, Crump, Johns, & Mewhort, 2020; Jones, 2018) store each sentence that the model experiences into memory. It would be easy to integrate social information into this framework by adding a signal of who produced a word into an instance. One could then form combined cue of words and person (e.g., *Jennifer* + *hockey*) to retrieve specific information

about one's linguistic experience with an individual (in this case, what you heard Jennifer say about hockey). Vector accumulation models, such as BEAGLE (Jones & Mewhort, 2007), could integrate social signals by also accumulating socially-based information (e.g., who produced a word in what context). Neural embedding models, such as word2vec (Mikolov et al., 2013) could have a separate input layer that signals communicative information for the model to learn. Other frameworks may need a greater level of adjustment to accommodate social information into their learning mechanisms, which is an important topic for future research.

An issue in distributional modeling that has not received much attention in distributional modeling but is a standard consideration in computational cognitive modeling (e.g., Shiffrin, Lee, & Kim, 2008), is that of model complexity. Distributional models differ in both computational complexity (Recchia & Jones, 2009) and parameter space (Johns et al., 2019). For example, word2vec has a large parameter space, and when other models (such as PMI; Bullinaria & Levy, 2007) are modified with a similar parameter space, they tend to have equal performance (Levy, Goldberg, & Dagan, 2015). In this article, concerns about model comparisons were overcome by using a single framework that integrated different types of information, meaning that the computational complexity and parameter space were identical for each model. However, going forward researchers need to be cognizant of these issues, especially as ever more sophisticated learning mechanisms are being proposed, to ensure that the development of distributional models of cognition cohere to standard practice in computational cognitive modeling and the development of cognitive theory.

Related to the modification of distributional models to accommodate social information is the need for training materials that contain social and communicative information. Here we used communication patterns from Reddit, made available from Baumgartner et al. (2020), following the previous work of Johns (2021). The use of social media and related sources have been used in previous studies, including Facebook (Park, Conway, & Chen, 2018; Schwartz et al., 2013), Twitter (Herdağdelen and Marelli, 2017), and Urban Dictionary (Johns, 2019). However, the popularity of these sources still pale in comparison to more classical training materials, such as textbooks (Landauer & Dumais, 1997), television and movie subtitles (Brysbaert & New, 2009), and online encyclopedias (Shaoul & Westbury, 2010b), among others. For future work in corpus construction, attention should be paid to demographic characteristics of who produced that language (see Johns & Jamieson, 2018, 2019; Taler, Johns, & Jones, 2020) in order to determine the impact of those characteristics on fits to different sets of human behavior (Johns, Jones, & Mewhort, 2019). As newer and better corpora are constructed, it is important to continue to integrate extra-linguistic information into a corpus in order to determine the types of information that can be used to construct better distributional models. Compared to integrating perceptually grounded information into a model's resulting representation, which requires other training material types (such as images or feature norms), the integration of social and communicative information into a corpus is relatively simpler, as it is a natural part of most language sources (e.g., who produced it, where it was produced, the date it was produced, the medium in which it occurred). However, corpora need to be constructed with this in mind.

The data used here to evaluate the different models on lexical semantics were mainly word similarity values. This type of data is likely biased towards the WW model, as this model directly measures the co-occurrence of two words in context. Future efforts should aim to collect socially-based data to determine how different distributional information is used across tasks. For example, distributional models have been used to assess biases in language and how they relate implicit biases in individuals (e.g., Caliskan, Bryson, & Narayanan, 2017). By taking into account the social and communicative aspects of model training materials, it may be possible to develop a better understanding on the spread of different forms of bias across a large number of people, especially on social media. Another possibility for socially-orientated distributional models is in sentiment analysis (e.g., Recchia & Louwerse, 2015) where distributional modeling is used to infer affective properties of text. It is likely that information about who produced a word, and in what discourse it was used in, could provide additional insight into inferring these properties.

The success of distributional models of semantics demonstrates the systematic connection between word co-occurrence statistics and the word meanings that people have acquired. However, the use of language is not separate from the context in which it occurs. The contexts that language is used in are multi-faceted and include social and communicative information. The results reported here demonstrate that a model that encompasses the communication patterns of how individuals use words across discourses allows for unique measures of word similarity to be derived, signifying the usefulness of integrating alternative information sources into distributional models of semantics.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by Natural Science and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2020-04727.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review, 23*, 275–329.
Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814–823.
Adelman, J. S., & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review, 115*(1), 214–227.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*(3), 463–498.
Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*, 283.
Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445–459.
Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences, 106*(41), 17284–17289.
Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 238–247).
Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(4), 577–660.
Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*(1), 617–645.
Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020, May). The pushshift reddit dataset. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 14, pp. 830-839).
Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., … Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning, 59*, 1–26.
Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review, 124*(1), 1–20.
Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modelling human judgment. *Current Opinion in Behavioral Sciences, 29*, 31–36.
Bhatia, S., & Stewart, N. (2018). Naturalistic multiattribute choice. *Cognition, 179*, 71–88.
Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*(3), 355–387.
Borghi, A. M., & Binkofski, F. (2014). *Words as social tools: An embodied view on abstract concepts* (Vol. 2). New York: Springer.
Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research, 49*, 1–47.
Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.
Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology, 7*, 1116.
Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods, 51*(2), 467–479.
Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review, 8*(3), 531–544.
Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39*(3), 510–526.
Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods, 44*(3), 890–907.
Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). Distributional semantics in technicolor. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (pp. 136-145). Association for Computational Linguistics.
Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186.
Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences, 31*(5), 489–509.
Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204–256.
Crump, M. J. C., Jamieson, R. K., Johns, B. T., & Jones, M. N. (2020). Controlling the retrieval of general versus specific semantic knowledge in the instance theory of semantic memory. *Proceedings of the 42nd Annual Cognitive Science Society*.
De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and Affective Multimodal Models of Word Meaning in Language and Mind. *Cognitive Science, 45*, Article e12922.
Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. In Proceedings of the 10th international conference on World Wide Web (pp. 406-414). ACM.
Firth, J. R. (1957). *Papers in linguistics 1934–51*. Oxford: Oxford University Press.
Gärdenfors, P. (2018). Levels of communication and lexical semantics. *Synthese, 195*(2), 549–569.
Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language, 43*(3), 379–401.
Goh, W. D., Yap, M. J., & Chee, Q. W. (2020). The Auditory English Lexicon Project: A multi-talker, multi-region psycholinguistic database of 10,170 spoken words and nonwords. *Behavior Research Methods, 52*(5), 2202–2231.
Goldberg, Y., Levy, O., 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722.
Goldinger, S. D. (1998). Echoes of echoes? An episodic trace theory of lexical access. *Psychological Review, 105*, 251–279.
Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244.
Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*(6), 1006–1033.
Halawi, G., Dror, G., Gabrilovich, E., & Koren, Y. (2012). Large-scale learning of word relatedness with constraints. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and Data Mining (pp. 1406-1414). ACM.
Herdağdelen, A., & Marelli, M. (2017). Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science, 41*(4), 976–995.
Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119*(2), 431–440.
Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411–428.
Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods, 45*(3), 718–730.
Hollis, G. (2017). Estimating the average need of semantic knowledge from distributional semantic models. *Memory and Cognition, 45*(8), 1350–1370.
Hollis, G. (2020). Delineating linguistic contexts, and the validity of context diversity as a measure of a word's contextual variability. *Journal of Memory and Language, 114*, 104146.
Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children's word reading. *Journal of Memory and Language, 103*, 114–126.
Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior, 1*(2), 119–136.
Johns, B. T., & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin & Review, 17*(5), 662–672.
Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology, 65*(4), 486–518.
Johns, B. T., & Jones, M. N. (2012). Perceptual Inference through global lexical similarity. *Topics in Cognitive Science, 4*, 103–112.
Johns, B. T., & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology, 69*(3), 233–251.
Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review, 23*(4), 1214–1220.
Johns, B. T., & Jones, M. N. (2021). Content matters: Measures of contextual diversity must consider semantic content. *PsyArXiv*.
Johns, B. T., Taler, V., Pisoni, D. B., Farlow, M. R., Hake, A. M., Kareken, D. A., et al. (2018). Cognitive modeling as an interface between brain and behavior: Measuring the semantic decline in mild cognitive impairment. *Canadian Journal of Experimental Psychology, 72*(2), 117–126.

Johns, B. T., & Jamieson, R. K. (2018). A large-scale analysis of variance in written language. *Cognitive Science, 42*(4), 1360–1374.
Johns, B. T., & Jamieson, R. K. (2019). The influence of time and place on lexical behavior: A distributional analysis. *Behavior Research Methods, 51*, 2438–2453.
Johns, B. T., Mewhort, D. J. K., & Jones, M. N. (2019). The role of negative information in distributional semantic learning. *Cognitive Science, 43*(5), e12730. https://doi.org/10.1111/cogs.2019.43.issue-510.1111/cogs.12730
Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review, 26*(1), 103–126.
Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2021). A continuous source reinstatement model of true and false recollection. *Canadian Journal of Experimental Psychology, 75*, 1–18.
Johns, B. T., Dye, M., & Jones, M. N. (2020). Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology, 73*(6), 841–855.
Johns, B. T., Jamieson, R. K., & Jones, M. N. (2020). The continued importance of theory: Lessons from big data approaches to cognition. In S. E. Woo, R. Proctor, & L. Tay (Eds.), *Big Data Methods for Psychological Research: New horizons and Challenges*. APA Books.
Johns, B. T., Jamieson, R. K., Crump, M. J. C., Jones, M. N., & Mewhort, D. J. K. (2020). Production without rules: Using an instance memory model to exploit structure in natural language. *Journal of Memory and Language, 115*, 104165. https://doi.org/10.1016/j.jml.2020.104165
Johns, B. T. (2019). Mining a crowdsourced dictionary to understand consistency and preference in word meanings. *Frontiers in Psychology, 10*, 268 (14 pages).
Johns, B. T. (2021). Disentangling contextual diversity: Communicative need as a lexical organizer. *Psychological Review, 128*(3), 525–557.
Johns, B. T. (in press). Accounting for item-level variance in recognition memory: Comparing word frequency and contextual diversity. Memory & Cognition.
Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1–37.
Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology, 66*(2), 115–124.
Jones, M. N., Willits, J., & Dennis, S. (2014). Models of semantic memory. In J. R. Busemeyer & J. T. Townsend (Eds.) Oxford Handbook of Mathematical and Computational Psychology.
Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an organizational principle of the lexicon. In B. Ross (Ed.), The Psychology of Learning and Motivation, 67:43.
Jones, M. N. (2018). When does abstraction occur in semantic memory: Insights from distributional models. *Language, Cognition and Neuroscience, 34*, 1338–1346.
Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44*(1), 287–304.
Kumar, A. A. (2020). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, 1–41.
Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review, 12*(4), 703–710.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211–240.
Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science, 41*, 677–705.
Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *In Advances in Neural Information Processing Systems* (pp. 2177–2185).
Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embedding. *Transactions of the Association for Computational Linguistics, 3*, 211–225.
Lieven, E. V., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language, 24*(1), 187–219.
Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods, 52*(3), 1271–1291.
Mak, M. H. C., Hsiao, Y., & Nation, K. (2021). Anchoring and contextual variation in the early stages of incidental word learning during reading. *Journal of Memory and Language, 118*, 104203. https://doi.org/10.1016/j.jml.2020.104203
Mandera, P., Keuleers, E., & Brysbaert, M. (2020). Recognition times for 62 thousand English words: Data from the English Crowdsourcing Project. *Behavior Research Methods, 52*(2), 741–760.
Marelli, M., & Baroni, M. (2015). Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological Review, 122*(3), 485–515.
Marelli, M., Gagné, C. L., & Spalding, T. L. (2017). Compounding as Abstract Operation in Semantic Space: Investigating relational effects through a large-scale, data-driven computational model. *Cognition, 166*, 207–224.
McRae, K., Cree, G. S., Seidenberg, M. S., & Mcnorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*(4), 547–559.
Mewhort, D. J. K., Shabahang, K. D., & Franklin, D. R. J. (2018). Release from PI: An analysis and a model. *Psychonomic Bulletin & Review, 25*(3), 932–950.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *In Advances in Neural Information Processing Systems* (pp. 3111–3119).
Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111*(3), 721–756.
Osth, A. F., Shabahang, K. D., Mewhort, D. J. K., & Heathcote, A. (2020). Global semantic similarity effects in recognition memory: Insights from BEAGLE representations and the diffusion decision model. *Journal of Memory and Language, 111*, 104071. https://doi.org/10.1016/j.jml.2019.104071
van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language, 50*, 289–335.
Park, A., Conway, M., & Chen, A. T. (2018). Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: A text mining and visualization approach. *Computers in Human Behavior, 78*, 98–112.
Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin, 134*(3), 427–459.
Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011, March). A word at a time: Computing word relatedness using temporal semantic analysis. In Proceedings of the 20th International Conference on World Wide Web (pp. 337-346). ACM.
Recchia, G. L., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information to latent semantic analysis. *Behavior Research Methods, 41*, 657–663.
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*(8), 1584–1598.
Recchia, G., Sahlgren, M., Kanerva, P., & Jones, M. N. (2015). Encoding sequential information in semantic space models: Comparing holographic reduced representation and random permutation. *Computational Intelligence and Neuroscience, 2015*, 1–18.
Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science, 3*, 303–345.
Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8*(10), 627–633.
Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE, 8*(9), e73791.
Shaoul, C., & Westbury, C. (2010a). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods, 42*(2), 393–413.
Shaoul, C., & Westbury, C. (2010b). *The Westbury Lab Wikipedia Corpus 201*. Edmonton, AB: University of Alberta.
Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32*, 1248–1284.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166.
Snefjella, B., Lana, N., & Kuperman, V. (2020). How emotion is learned: Semantic learning of novel words in emotional contexts. *Journal of Memory and Language, 115*, 104171. https://doi.org/10.1016/j.jml.2020.104171
Taler, V., Johns, B. T., Young, K., Sheppard, C., & Jones, M. N. (2013). A computational analysis of semantic structure in bilingual fluency. *Journal of Memory and Language, 69*, 607–618.

Taler, V., Johns, B. T., & Jones, M. N. (2020). A large scale semantic analysis of verbal fluency across the aging spectrum: Data from the Canadian longitudinal study on aging. *Journal of Gerontology: Psychological Sciences, 75*, e221–e223.
Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.
Tomasello, M. (2009). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
Westbury, C., & Hollis, G. (2019). Conceptualizing syntactic categories as semantic categories: Unifying part-of-speech identification and semantics using co-occurrence vector averaging. *Behavior Research Methods, 51*(3), 1371–1398.
Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review, 18*(4), 742–750.