Cognitive Systems Research 80 (2023) 90–102

Contents lists available at ScienceDirect

Cognitive Systems Research

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

Computing word meanings by aggregating individualized distributional models: Wisdom of the crowds in lexical semantic memory

Brendan T. Johns *

McGill University, Canada

ARTICLE INFO

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

ABSTRACT

Linguistic experience varies across individuals and is impacted by both demography and personal preferences, leading to differences in word meanings across languages (Thompson et al., 2020) and people (Johns, 2022). An active area of study in the cognitive sciences that examines the impact of varied knowledge across individuals is the wisdom of the crowd effect, where it is found that the aggregate judgement of a group of individuals is often better than the judgement of the best individual in the group (Surowiecki, 2004). The goal of this article was to determine if there is a wisdom of the crowd effect in lexical semantic memory, such that the aggregated word similarity values from many individual language users exceeds the fit of the best fitting individual. This was accomplished by training 500 different distributional models from 500 high-level commenters on the internet forum Reddit. By deriving aggregated word similarity values from these individuals, a strong wisdom of the crowd effect was found where the aggregated similarity values far exceeded the performance of the best fitting individual for each dataset tested. Additionally, it was found that even aggregating only a small number of users provided a large increase in fit relative to the individual corpora, but with the best fitting measure including word similarity values from all possible users. The results of this article provide an avenue for future distributional model development by demonstrating that the best pathway towards better distributional models may lie in the aggregation of multiple representations attained from individual users of a language.

1. Introduction

In a classic experiment on reconstructive memory, Bartlett (1928, 1932; see also Bergman & Roediger, 1999 for a modern replication) had participants read an Indigenous American story entitled "The War of the Ghosts." When asked to recall the story, British participants inserted their own accumulated knowledge into their memories of the narrative. For example, instead of using the word *canoe* (a word they were not familiar with) in their recollection of the story, participants tended to use the word *boat* instead. This finding suggests that language comprehension and usage is not independent of an individual's unique experience, as the concepts that one has acquired through that experience are used in the understanding and recall of a new context. To comprehend a current linguistic context requires the utilization of linguistic representations derived from past experiences with language, with that experience varying appreciably across people based on culture and demography.

The primacy of the role of linguistic experience on language comprehension has been most aptly evaluated in computational cognitive modeling, namely through the development and use of distributional models of lexical semantics. There are multiple models of this type (e.g., Griffiths et al., 2007; Jamieson, Avery, Johns, & Jones, 2018; Jones & Mewhort, 2007; Landauer & Dumais, 1997; Mikolov et al., 2013a) and all are based around the premise that a word's meaning can be constructed through an observation of how a word is used across lexical contexts (e.g., sentences, documents), specifically through the learning of word co-occurrence statistics. This model type has had substantial success at explaining a variety of different behaviors across the study of language and memory (for recent reviews, see Bhatia & Aka, 2022; Günther et al., 2019; Kumar, 2020).

Distributional models require large corpora of natural language to derive word meanings from. That is, this model type is experientially dependent, and the resulting behavior of the model is impacted by the structure of the training materials that the model is learning from. There are a number of different corpus types that have been employed over the years to train distributional models and all differ substantially in their underlying source materials, such as online encyclopedias (Shaoul & Westbury, 2010), textbooks (Landauer & Dumais, 1997), television and

* Address: Department of Psychology, McGill University, 2001 McGill College Avenue, Montreal, Quebec H3A 1G1, Canada.
E-mail address: brendan.johns@mcgill.ca.

https://doi.org/10.1016/j.cogsys.2023.02.009
Received 2 December 2022; Received in revised form 15 February 2023; Accepted 26 February 2023
Available online 5 March 2023

movie subtitles (Mandera, Keuleers, & Brysbaert, 2017), fiction books (Johns et al., 2020; Johns & Jamieson, 2018, 2019), social media (Herdağdelen and Marelli, 2017; Johns, 2019, 2021a,b), and newspaper articles (Aujla, 2021; Davies, 2009), among others. Most modern corpora used to train distributional models lie in the hundreds of millions to billions of words of text and the best corpus to explain a linguistic behavior differs appreciably across the tasks being examined (Johns et al., 2019).

However, text corpora should be appreciated for what they are: samples of language that are typically written by many authors towards a targeted audience. That is, a large corpus contains an accumulation of written language from many different individuals, written for different purposes and at different times. The word meanings that are derived from these text sources are thus an average of many different people's usage of language across different types of written language. However, the experience that a specific individual has with language is determined by a multitude of factors, such as one's age, socio-economic status, educational background, political alignment, and entertainment preferences, among many other factors. Thus, the linguistic experience that any single individual has is likely considerably different from a corpus that attempts to encode what average linguistic experience is (e.g., it is unlikely that any single individual has read all of Wikipedia).

Johns and Jamieson (2018) provided an initial examination into individual differences in language usage at a large scale by examining the distributional properties of books written by popular fiction authors. It was found that books written in the same genre (e.g., romance novels) had similar distributional properties when compared to books written in different genres. However, this difference in similarity was dwarfed in comparison to what was found at the single author level, where it was shown that books written by an individual author were much more similar to each other than books written by different authors, regardless of the genre the books were written in. This effect was referred to as the author signature effect, as it demonstrates that each individual author has a unique signal of how they use language (see Johns, Dye, & Jones, 2020 for a replication of this effect with a much larger set of books). Johns and Jamieson (2019) followed up on this work by demonstrating that lexical organization and lexical semantic data collected from the UK and USA, and at different timepoints, were best explained by time- and place-appropriate text corpora (e.g., the familiarity data of subjects from the UK was best fit by word frequency values from a corpus of books written by authors born in the UK, and vice-versa for data collected in the USA). Johns, Jones, and Mewhort (2019) provided a framework to optimize experience-based cognitive models by automatically manipulating the type of text that is used to train a model, and it was found that the text selected is coherent with the experiential properties of the group that the behavioral data was collected from (e.g., the optimization procedure selected young adult fiction novels to explain data collected from young adults).

Combined, the results of Johns and Jamieson (2018, 2019) and Johns et al. (2019) suggest that language usage differs substantially at the individual level. This finding is strengthened by recent results by Thompson, Roberts, and Lupyan (2020), who employed distributional models to examine how relative word meanings are across different languages. Thompson et al. (2020) found that word meanings differed radically across languages and that these differences were, at least in part, explained by cultural differences in the countries where the languages were used. Johns (2022) followed up on this work by constructing corpora for 500 different commenters on the internet forum Reddit, based upon previous work by Johns (2021a,b; Johns & Jones, 2022) using Reddit to examine social and communicative factors in lexical organization and lexical semantics. Johns (2022) found that there were significant levels of variability for word meanings across different individuals, suggesting that word meanings differ appreciably across members of a language speaking community.

The results of Thompson et al. (2020) and Johns (2022) begs the question as to what the inherent variability in words meanings at both the cultural and individual level has on the theoretical understanding of lexical semantic memory. This is a particularly poignant concern for distributional models of semantics as these models are trained on average, not individual, levels of language experience. However, there is an area study in both cognitive and social psychology that examines the impact of varied knowledge across individuals, namely the study of the wisdom of the crowd effect.

Wisdom of the crowd refers to the finding that the aggregate judgement of a group of individuals is often better than the judgement of the best individual in the group (see Surowiecki, 2004; Steyvers & Miller, 2015; Yaniv, 2004). This effect was most famously demonstrated by Galton (1907) who had 800 spectators provide subjective estimates for the weight of an ox and found that the average of the estimates was only one pound off from the correct weight (see also Gordon, 1924 for early empirical results on the accuracy of groups over individuals). A wisdom of the crowd effect has been found across a wide variety of problem types, such as in recollection of ordered information (Steyvers et al., 2009), complex combinatorial problems (Yi, Steyvers, Lee, & Dry, 2012), and forecasting (Armstrong, 2001; Mannes, Soll, & Larrick, 2014; Merkle, Steyvers, Mellers, & Tetlock, 2017). Determining the optimal combination of individual judgements has also been the target for computational cognitive modeling (e.g., Lee & Danileiko, 2014; Lee, Steyvers, & Miller, 2014; Lee, Zhang, & Shi, 2011).

In terms of lexical semantic memory, the question that this article seeks to understand is whether the aggregate word meanings across individuals are more accurate than the word meanings from the best fitting individual. This is a slightly different interpretation of the wisdom of the crowd effect than previous work, as typically this effect is examined in the context of aggregating judgements across individuals (e.g., asking 100 economists to forecast the state of the economy in 6 months and using the averaged judgement of the economists as a prediction). Instead, here the wisdom of the crowds will be interpreted in terms of whether the averaged similarity values from the individual user models are closer to commonly accepted word meanings (as operationalized from the similarity values given by human participants in behavioral experiments) than the individual models themselves or a model trained on all corpora concurrently. In contrast to the aggregation of judgements across people, where there is an objective reality that the judgements can be compared against (e.g., the state of the economy 6 months after the predictions were made), examining word meanings lacks this objectivity.

Instead, the examination wisdom of the crowd effect as investigated in this article is more akin to a model testing framework, where different approaches to aggregating individual knowledge will be contrasted to determine the best way to capture word similarity judgements. If it is found that the best fitting model is one that aggregates individual similarity values across multiple models, it would suggest that in order to capture population-level knowledge about the meaning of words a model is better served by integrating individual differences into its resulting similarity judgements, rather than incorporating as much language as possible. However, given that distributional models are now being used in cognitive models of judgement and decision (e.g., Bhatia, Richie, & Zou, 2019; Singh, Richie, & Bhatia, 2022), the results of this article provide an existence proof that using multiple individualized distributional models could serve as a solid grounding for developing more accurate models of judgement and decision.

An additional goal of this article is to determine the shape of the improvement in fit across individuals, specifically what the benefit in model fit is as a greater number of word meaning estimations from different individuals are included in the aggregate word similarity judgements. The theoretical motivation for this work is the recognition of the inherent variability in the linguistic experience that different people have and the resulting impact on word meaning representations that are derived from this experience. If aggregating word similarity values across individuals shows a large increase in performance (i.e., a wisdom of the crowd effect is found), it would suggest that distributional

models should focus on accounting for knowledge acquisition at the individual level and average word meanings should be determined with aggregation techniques.

The determination of a wisdom of the crowd effect in lexical semantic memory will be accomplished by the assembling of 500 individual user corpora from high-level commenters on the internet forum Reddit, following the work of Johns (2022). A distributional representation, based on a count-based methodology (Johns, Mewhort, & Jones, 2019; Johns, 2021b, 2022), will be derived for each user corpus and different methods of aggregating the word similarity values from these representations will be evaluated.

To provide an example of the variability in word meanings across individual language users contained in these user corpora, Table 1 displays the nearest neighbors (the most 15 similar words) to the word *freedom* for eight different individual corpora using the modeling framework described below. As can be seen from the table, all users generate words that are related to the term freedom (e.g., *religion*, *violate*, *sexual*, *privacy*), but most of the words generated are different across the users. There are some consistent words in the neighborhoods – for example, the word *liberty* is a near neighbor for 6 out of 8 of the users. However, many of the terms are unique to each individual, suggesting that every individual user has a distinctive representation of the meaning of the word *freedom*. In terms of a wisdom of the crowd effect, this demonstration shows that the word meanings derived from different individuals vary appreciably. The question becomes whether aggregating across these different word similarity values leads to better fits to lexical semantic data compared to any single individual user of language.

Distributional models have furthered cognitive theory by demonstrating the systematic connection between the lexical experience that people have with the natural language environment and lexical semantic behaviors. Standard practice in distributional model training is to employ a corpus that contains a very large amount of natural language, in order to certify that the model has enough experience with all wordforms, such that they can be ensured to be forming reliable representations of those words. However, this is not coherent with the lexical experience of the people that these models seek to simulate. Instead, the lexical experience that people have is directed and impacted by cultural and demographic properties of their life. By examining the output of multiple distributional models, trained on the language usage of specific individuals, this article will provide theoretical insight into how distributional models combine at the group level to form accurate judgements of word meanings. Additionally, the results of this article will provide an avenue for future distributional model development by demonstrating that the best pathway towards better distributional models may not lie in the maximization of training materials for a proposed model, but instead through the aggregation of multiple representations attained from individual users of a language.

2. Methods

Four aspects of the modeling work will be described here: (a) the individual user corpora, (b) the distributional modeling framework, (c) the word similarity aggregation methods, and (d) the behavioral data that will be used to evaluate the individual and aggregated models.

2.1. Individual user corpora

Individual corpora for 500 users attained from the internet forum Reddit were used here, with these corpora having previously been used by Johns (2022; much larger sets of user corpora were previously used in Johns, 2021a,b to determine the impact of social and communicative information on lexical organization and lexical semantics). The Reddit corpora were constructed from publicly available database files on the website pushshift.io (Baumgartner, Zannettou, Keegan, Squire, & Blackburn, 2020), where all Reddit comments are posted as database files for each month, built through the publicly available Reddit API. All comments made on Reddit are organized by month, with each month being posted as a JSON file that contains the text of each comment and associated metadata (such as the user who produced the comment). Corpora for the 500 most prolific commenters on Reddit were constructed by building individual corpora for all commenters (who have a publicly available usernames) who produced more than 3,000 comments on the site and selecting the 500 who produced the greatest amount of language (not necessarily the greatest number of comments). All comments through to July 2021 were used to construct each individual user corpus.

Each corpus was hand-inspected to determine that they consisted of real comments and not those produced by an automated bot. Additionally, a criterion was set such that each corpus had to contain at least 10,000 word types, ensuring that each individual corpus contains a variety of different wordforms. As described in Johns (2022), each corpus has on average 7.3 million words, with the largest individual corpus consisting of 32.3-million-word tokens and the smallest consisting of 3.8-million-word tokens. The total number of words across all corpora was 3.8-billion-word tokens. The average number of word types was 30,008, with the corpus with largest number of types having 52,032 and the lowest having 16,996. The number of types for a user roughly maps onto the productive vocabulary of an individual and is likely impacted by the type of discourses a user communicates within, where discussion of more technical topics (e.g., computer programming) leads to a greater vocabulary size. See Johns (2022) for a more in-depth analysis of the content of these corpora and discussions of quality control.

Table 1
Ranked nearest neighbors for 8 different users to the word “freedom.”.

User_1	User_2	User_3	User_4	User_5	User_6	User_7	User_8
Religion	Patriot	Resistance	Sexual	Speech	Hypocritical	Responsibility	Privacy
Liberty	USA	Protected	Form	Protecting	Hating	Greed	Violate
Rights	Privileges	Hating	Power	Idea	Atheist	Liberty	Protects
Expression	Creativity	Free	Blast	Free	Terrorist	Imagined	Liberty
Absolute	Strive	Liberty	Functioning	Belong	Hungry	Cheat	Citizens
Mobility	Pilot	French	Activity	Privacy	Radical	Ethics	Celebrating
Violation	Rights	Monarchy	Relaxation	Defending	Potatoes	Europeans	Compromise
Institution	Beliefs	Meeting	Desire	Loving	Socialist	Assumes	Democracy
Assembly	Liberty	Proposed	Crowd	Crowded	Blaming	Expose	Nation
Economics	Priests	Loving	Surge	Restricted	Greedy	Definition	Survival
Artistic	Act	Dignity	Strengths	Factor	Fool	Christianity	Intuitive
Promise	Lecture	Amendment	Vein	Religion	Communist	Political	Protesting
Slavery	French	Speech	Genders	Allowing	Speech	Capitalism	Borders
Torture	Unity	Tolerance	Nature	Threats	Advocating	Economic	Hate
Constitution	Free	Healthcare	Release	Liberty	Scandal	Commitment	Invade

2.2. Distributional modeling framework

There are currently many different distributional models used in current research (see Kumar, 2020 for a review). Here, a count-based distributional model developed by Johns, Jones, & Mewhort (2019) and previously used by Johns (2021b, 2022) will be utilized. This model was developed in order to elucidate the role that the optimization methods that are employed by neural embedding distributional models (e.g., Word2Vec; Mikolov et al., 2013; Mandera et al., 2017) play in accounting for word similarity data. Neural embedding models use a predictive neural network to generate predictions about the words that should surround a target word in context and use an error signal to improve a word's ability to predict the words that should co-occur with that word. Neural embedding models also use a procedure entitled negative sampling, where a number of unrelated words are generated based upon word frequency and the network is made less predictive of these unrelated words. Johns et al. (2019) demonstrated that negative sampling does not serve to hone a prediction method, but instead serves to integrate base-rate co-occurrence information into the model's representation allowing for important word-word associations to be highlighted. Furthermore, Johns et al. (2019) demonstrated that by integrating this information into a simpler, count-based representation allowed for equivalent fits to neural embedding models to be attained while vastly reducing the number of parameters in the model.

Count-based distributional models are simple models that accumulate word-word co-occurrences across a corpus, with the most common version of this model type being pointwise mutual information (PMI; Bullinaria & Levy, 2007, 2012), an information theoretic metric defining the probability of two words co-occurring together in context. PMI has been demonstrated to provide similar fits to neural embedding models, despite being considerably less complex (Levy & Goldberg, 2014; Levy, Goldberg, & Dagan, 2015). As detailed in Johns (2022), count-based representations offer a number of advantages when exploring the use of small corpora. The most notable advantage is that neural embedding models have been demonstrated to struggle when trained on small corpora (Levy, Goldberg, & Dagan, 2015), likely due to the model's utilization of probabilistic functions based on precomputed word frequency distributions, such as negative sampling and subsampling (the probabilistic skipping of high frequency words). Word frequency distributions from small corpora may not be consistent enough to construct reliable probabilistic functions.

The underlying representation of the distributional modeling framework used here is a Word × Feature matrix, where each row is a word's representation and each column is a feature of some type, with the feature used here being other words (see Johns, 2019, 2021b for exploration of the framework with other feature types). Thus, here the dimensionality of the matrix is $W \times W$, where W is the size of the model's vocabulary. Each element in the matrix is the number of times each word co-occurred together within a window in a sentence. Although window size is technically a free parameter, here it will be simply set as co-occurrence within an entire sentence, in order to simplify the model by removing a free parameter and eliminate the need to optimize the model across users. This will be referred to as the WW model.

Although the untransformed matrix provides a poor accounting of word similarity data, Johns et al. (2019) developed two parameter-free matrix transformations that allowed for the model to form accurate semantic representations. These transformations were developed based upon the finding that integrating negative sampling into the WW model's training routine allowed for a massive increase in model performance due to it allowing for unique word-word co-occurrence values to be highlighted. If two words occurred together more than the probability of a word being randomly sampled, that co-occurrence signals a unique association for that word. The matrix transformations developed Johns et al. (2019) allowed for the WW model to achieve similar levels of fits to neural embedding models, while being much simpler (see also Shabahang, Yim, & Dennis, 2022 for additional evidence about the importance of negative information in lexical semantics). A further advantage for the goals of this article is that the model does not employ any probabilistic functions, which means each of the 500 different user models will be trained with the exact same model setup.

Three transformations are used to hone the representation formed by the WW model: 1) global negative (GN), 2) distribution of associations (DOA), and 3) combined GN + DOA. The GN transformation seeks to directly integrate base-rate co-occurrence information into the model's representation and is a direct analogue to the negative sampling operation used by neural embedding models. The DOA transformation attempts to highlight unique connections based upon the latent pattern of co-occurrences across the matrix and the combined transformation exploits the fact that the GN and DOA transformations are tapping into slightly different information.

The first step in applying the GN transformation is to construct a global negative vector, which contains the base-rate occurrence values across the columns of the matrix. To compute this vector, the sum of each column in the matrix is calculated:

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

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 elements in the GN vector is dependent on the frequency of the word across sentences. The second step of the transformation is to unit normalize the vector to have a total magnitude of 1 by dividing each element by the total sum of the elements in the vector:

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

where k increments through each index in the **GN** vector with the prime mark indicating the transformed value.

The third and final step of the transformation is to balance the amount of positive and negative information in a word's representation in the matrix. This is done by computing the sum of a word's row (i.e., the total number of co-occurrences that have been accumulated during a word's training) and adding in an equal amount of negative information:

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

where $\mathbf{M}_i$ is a word's row in the matrix, and j goes through each column in the matrix. The resulting transformed matrix has an equivalent amount of negative and positive information contained in a word's representation. A positive value in the matrix signals that the connection between two words exceeds the base-rate occurrences of those words, while if it is negative or close to zero it signals that the words do not have a unique connection to each other.

Unlike the GN transformation where the levels of positive and negative information is exactly balanced, the DOA transformation uses the fact that the uniqueness of two words co-occurrence patterns is contained in the matrix through the relative magnitude of the elements in the matrix. To capitalize on this realization, the first step in the DOA transformation is to transform the columns of the matrix into 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 values in the resulting matrix signal how many standard deviations the association between two words is in relation to the other values in the column. This signals how unique a co-occurrence value is.

However, the resulting z-scores are biased by the word frequency of a word, where higher frequency words will have high z-scores on average

just because they tend to occur in more sentences. Thus, the second step is to transform each row into a z-score, in order to emphasize the important associations within a single word:

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

where μ_i is the mean of a word's row, and σ_i is the standard deviation of that row. The result of this transformation is that a word's representations contains normalized association values.

In Johns et al. (2019) and Johns (2021b) it was established that a combination of the GN and DOA transformations, where the GN transformation is applied first followed by the DOA transformation, provides the best fit to word similarity data. Importantly, Johns (2022) demonstrated that this same pattern holds for the smaller corpora used here, but with the DOA contributing the bulk of the increase in fit, but with the combined transformation still providing an increase in model performance. Thus, the combined transformation will be the one used here. One of the advantages of this approach is that it does not require the use of a stop list or subsampling to reduce the impact of high frequency words on semantic representations, as the contribution of these words is reduced through the transformations. This means that the resulting model only has a single parameter, that of vocabulary size. The vocabulary size used here will be the most frequent 30,000 words across the 500 user corpora, which means that the WW model will have a dimensionality of 30,000 × 30,000. The simplicity of the approach is advantageous for the current study as it allows for semantic representations of different corpora to be encapsulated with identical models, with only the content of the different corpora driving differences in the resulting user models.

The metric derived from the model to fit to behavioral dataset will be word pair similarity. A vector cosine is a normalized dot product and is computed with the following formula:

$$S(\mathbf{x}, \mathbf{y}) = \frac{\sum_{j=1}^{N} \mathbf{x}_j \times \mathbf{y}_j}{\sqrt{\sum_{j=1}^{N} \mathbf{x}_j^2} \sqrt{\sum_{j=1}^{N} \mathbf{y}_j^2}} \tag{6}$$

Where × and **y** are two vectors and *N* is the size of the vectors. The returned similarity value has a continuous range between −1 and 1, with 1 signalling complete overlap in vector features and −1 signalling opposite feature values across the vectors.

2.3. Aggregation techniques

The goal of this article is to compare the performance of the individual user corpora against the aggregated word similarity values across all users. However, there are multiple ways of aggregating the user corpora together. The first option is to simply train the WW model on all user corpora at once, so that the representation formed by the model will contain all the text produced by the 500 users. This will be referred to as Total Representation (TR) method and is consistent with standard techniques used in distributional modeling where the amount of text that a distributional model receives in training is maximized.

The second aggregation method is to instead train individual models for each user corpus, resulting in 500 different representations. Average word similarity between words is then computed by taking the similarity between those two words for each user representation and averaging the similarity values. This will be referred to as the Average Similarity of Words (ASW) method. However, one issue with this method, as previously explored in Johns (2022), is that since the individual user corpora are relatively small in size, some words may not be represented well in some of the user models compared to others. To overcome this issue, a frequency parameter will be used, where only words that exceed the parameter will be used to calculate average word similarity. Johns (2022) found that representations were relatively stable after 25 occurrences of a word, so in this study the frequency parameter will be set at 25 as well.

Although the TR and ASW methods will be trained the same amount of linguistic information, they have different theoretical bases. The TR method proposes that the best way to construct word meanings is to maximize the amount of training materials that a model receives, with the resulting representations being the average word meanings across all individual users with the unique language usage of each individual being removed. The ASW method takes into account individual differences in word meanings as the 500 user models will all have different representations for the same word. Thus, in the ASW method individual variability will have an impact on the resulting average word similarity values. The impact of this variability will be a central point of focus in the coming simulations.

2.4. Datasets

Three word similarity and one free association dataset will be used to evaluate the representations of the individual corpora and the different aggregation techniques. The three word similarity datasets are: (a) the WordSim data (n = 353; Finkelstein, et al., 2001), (b) the MTURK-771 data (n = 771; Halawi, Dror, Gabrilovich, & Koren, 2012), and (c) the MEN data (n = 3,000; Bruni, Boleda, Baroni, & Tran, 2012). The free association values are forward association strength (FAS) attained from the classic Nelson, McEvoy, and Schreiber (2004) norms. There were 49,995 number of pairs tested for the FAS data. Although distributional tend to not provide an overly strong correlation to free association data (Maki, 2008), due to the greater complexity of the task compared to word similarity, the number of pairs for the FAS data is much greater, which will allow for a better understanding of the performance of the different models with a larger selection of language.

The utilization of word similarity data to evaluate distributional models is standard in the field (e.g., Levy, Goldberg, & Dagan, 2015; De Deyne, Perfors, & Navarro, 2016). These data are collected by asking participants to rate the similarity of words on a scale, with some experiments manipulating task instructions. The data that models are fit to are the averaged word similarity values across participants. Thus, the data being fit to in this article are collapsed ratings that have underlying variance due to differences in word knowledge across the participants in the study. This is somewhat similar to the modeling approach utilized by the ASW method, where word similarity values are computed by averaging the similarity values from distributional models trained on individualized corpora. If it is found that that the ASW method provides superior performance compared to the TR method, it would suggest that variance in word meanings across language users needs to be considered when building models of lexical semantics.

2.5. Data and code availability

Code for the below simulations is available at https://osf.io/nbc6x/?view_only=e83de7fa221745b48a37846bd6c3d40c. Due to the size of the individual corpora and representations derived from them, as well as privacy concerns, the code builds ASW similarity values from precomputed word similarity values for each individual user corpora. However, the individual corpora are available upon request.

3. Results

The first analysis will compare the fit of the TR and ASW methods to the various word similarity datasets, as determining which method is best will set the course for the following analyses. For a word pair to be included for a dataset there had to be at least 30 individuals who had similarity values for that pair. This criterion was set so that there are stable average values for each word pair for the ASW method. Fig. 1 contains the Spearman correlations for both methods across the four different datasets. This figure shows that for all datasets the ASW method outperforms the TR method, suggesting that this is the superior

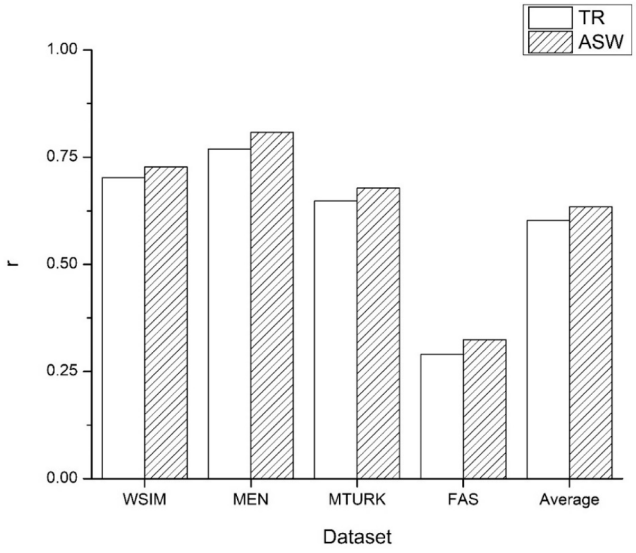

Fig. 1. The correlations for the various word similarity datasets to the ASW and TR methods of aggregating the individual user corpora. N = 275 for the WSIM data, N = 1,958 for the MEN data, N = 644 for the MTURK data, and N = 36,129 for the FAS data.

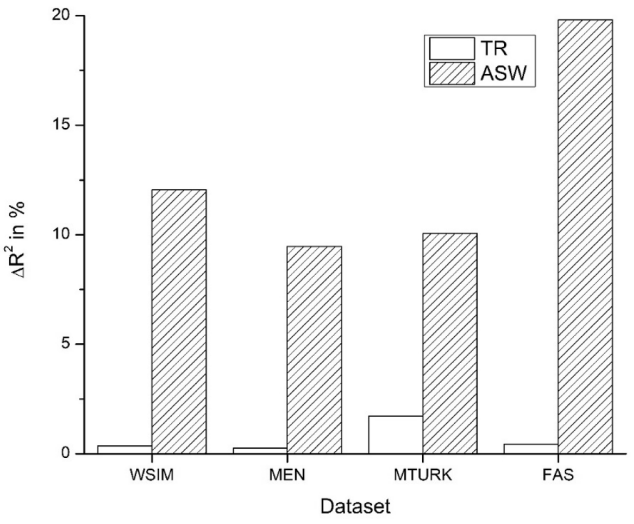

Fig. 2. Results of a hierarchical regression analysis comparing the amount of unique variance that the ASW and TR methods account for in the different word similarity datasets. All effects are significant at the $p < 0.01$ level, except for the TR method for the WSIM dataset which was not significant. N = 275 for the WSIM data, N = 1,958 for the MEN data, N = 644 for the MTURK data, and N = 36,129 for the FAS data.

method of aggregating lexical information across individual users. Additionally, the model fits displayed in Fig. 1 to the data are quite high and comparable to other models of this data (see Johns, Mewhort, & Jones, 2019 for model comparison values), suggesting that the underlying framework of the ASW method used here is a realistic model of lexical semantics. This is an intriguing finding as it suggests that aggregating the similarity values across multiple individual corpora outperforms a single model that is trained on all the corpora at once.

In order to directly compare the amount of variance explained by the ASW and TR methods for the different word similarity datasets, a hierarchical regression analysis was conducted. 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) and has been recently used by Johns (2021b) examining word similarity data. A hierarchical linear regression allows for one to assess the amount of predictive gain (measured as percent ΔR^2 improvement) for one predictor over other competing predictors, when they are contained in a linear regression. Separate regressions were run for each word similarity dataset, and Fig. 2 displays the amount of unique variance that the ASW and TR similarity values explain when compared against each other. This figure shows that for each dataset the ASW similarity values account for more variance than the TR values, while reducing or eliminating the variance accounted for by the TR values. This analysis conclusively demonstrates that the ASW method is providing better fits to the word similarity data as compared to the TR method.

One final concern about the analyses displayed in Figs. 1 and 2 is the impact that the number of users included in the averaged word similarity values has on ASW model performance relative to the TR model. This is an important consideration because the minimum number of users included in calculating the ASW word similarity values impacts the number of word pairs that are evaluated for each dataset. Thus, the TR model could be outperforming the ASW model if a different number of minimum users were evaluated. To test this possibility, an additional simulation was done where the number of minimum users in the ASW model was manipulated at 1, 10, 20, 30, 40, and 50 minimum users. Corresponding Spearman correlations and hierarchical regressions were used to compare the fits of the ASW and TR models across these levels, consistent with the previous simulations, as well the sample sizes utilized for each comparison level and dataset. The result of this simulation is contained in Table 2. This table shows that for all levels of minimum

Table 2
Impact of number of users and sample size on model fits and comparisons.

			r		ΔR^2 in %	
Dataset	# of Users	n	ASW	TR	ASW	TR
WSIM	1	348	0.7	0.67	11.462	3.162
	10	323	0.73	0.676	14.419	0.374
	20	296	0.726	0.68	12.5	.378*n.s.*
	30	275	0.729	0.702	12.05	.353n.s.
	40	264	0.731	0.692	11.02	.561n.s.
	50	261	0.727	0.687	11.32	.431n.s.
MTURK	1	764	0.665	0.647	8.932	3.703
	10	719	0.673	0.651	9.012	3.004
	20	678	0.674	0.648	9.892	2.15
	30	644	0.678	0.648	10.064	1.713
	40	607	0.68	0.65	10.021	1.705
	50	586	0.683	0.651	10.169	1.271
MEN	1	2,905	0.759	0.755	5.463	4.635
	10	2,402	0.793	0.765	8.176	1.1
	20	2,147	0.802	0.767	8.992	0.465
	30	1,958	0.809	0.771	9.465	0.252
	40	1,805	0.809	0.772	9.021	.153n.s.
	50	1,675	0.815	0.778	8.872	.15n.s.
FAS	1	48,482	0.314	0.296	14.563	4.854
	10	42,261	0.323	0.294	17.924	0.943
	20	38,595	0.325	0.292	19.626	0.647
	30	36,129	0.324	0.29	19.811	0.437
	40	34,522	0.325	0.29	20.754	0.372
	50	33,170	0.325	0.289	20.849	0.285

Note. *n.s.* = not significant.

users across all datasets the ASW model offered stronger correlations and explained more unique variance than the TR model. At a minimum user number of 1, the advantage of the ASW approach was decreased, but there was a sizeable advantage at only a minimum user inclusion level of 10. Combined with the results of Figs. 1 and 2, this simulation demonstrates that the ASW offers a better accounting of word similarity data than the more traditional TR approach to distributional semantics. The impact of number of users on ASW model performance will be a primary examination point explored in subsequent simulations.

The superiority of the ASW method over the more traditional distributional modeling approach as embodied by the TR method, suggests that in order to produce accurate estimations of word meanings

requires taking into account individual variability in the word meanings that people have. There is underlying variability in the behavior of people in psycholinguistic experiments on lexical semantics (e.g., asking people to rate the meaning of two words on a scale) and the increased power offered by the ASW method demonstrates that a distributional modeling approach that takes into account individual variability offers better fits to this type of data. An additional advantage of using the ASW approach is that it provides more flexibility in aggregating user similarity values, which will be exploited below to better understand the wisdom of the crowd effect in lexical semantics.

Before comparing the aggregated word similarity values to the fit of the individual user models, it needs to be established just how much variance there is in the underlying word representations that the method is utilizing. To accomplish this, a simulation was conducted comparing the word similarity values of each individual user to each other. To do this, the correlation between the word similarity values for each was pair was computed. For each user pair, the word pairs that the two users had in common (i.e., the word pairs where both users exceeded the frequency parameter) were attained. The possible word pairs included all pairs across the four datasets examined here (for a total maximum comparison of 55,031 word pairs). Then, the Pearson correlation coefficient was calculated for these values. This comparison was done for each possible user pair, resulting in 124,750 correlations. The result of this simulation is contained in Fig. 3, which displays the histogram of these correlations. The average correlation across user pairs was 0.427. The minimum correlation was 0.177, while the maximum correlation was 0.7, indicating a large spread in the user's semantic representations. This simulation demonstrates that the user corpora all contain substantially different word meaning representations, indicating that the similarity values that the ASW is using to compute averaged similarity values have significant variability across individual users.

The next point of knowledge that needs to be determined is how the fit of each model trained on the individual user corpora compares to the fit of the ASW values. Due to the use of the frequency parameter, each individual user will be tested with different word pair sets across the datasets. For each individual user, the pairs used will be the ones where both words exceed the frequency parameter for that specific individual. Thus, when computing the fit for the averaged similarity values to the individual models, only the word pairs that the individual was tested on were used for the ASW method. This means that there will be 500 different correlations for each of the individual and ASW methods. As stated previously, the utilization of the frequency parameter is an attempt to ensure that there is control over levels of linguistic information contained in the different models.

The result of this simulation is contained in Fig. 4. This figure shows that while the individual models do provide reasonable fits to the word similarity data, the ASW method vastly outperforms the individual user models. This finding suggests that taking into account the similarity values derived from diverse semantic representations across users outperforms a single individual's derived representation and demonstrates a powerful wisdom of the crowd effect in lexical semantic memory, as the aggregate similarity values significantly outperform the individual corpora.

However, Fig. 4 displays the averaged correlations across individuals and as the error bars show there is significant variation in the fit to the data across individuals. To truly be a wisdom of the crowd effect, the ASW method should exceed the best fitting individual and not just the average fit of the individual users. To ensure that this is the case, for each dataset the best fitting individual was found and compared to the ASW model for that individual's selected word pairs. The results are displayed in Fig. 5 and shows that for each dataset the performance of the ASW method exceeds the performance of the best fitting individual. Indeed, across all 2,000 comparisons (500 users across 4 different datasets) between an individual user corpus and the ASW method, the ASW method had the better fit. The results of Figs. 4 and 5 provide conclusive evidence of a wisdom of the crowd effect in lexical semantic memory, as the averaged values across individuals outperform the best fitting individual across every dataset. The following two simulations will attempt to clarify the underlying reasons for the success of the ASW approach.

The results contained in Fig. 4 demonstrate that when the ASW method accumulates similarity values from all users the model vastly outperforms the individual corpora. A secondary question is what impact the number of users included in the ASW similarity values has on model performance. To examine this, a simulation was conducted comparing the individual corpora fit to the ASW method when 2, 4, 6, 8, and 10 users were included in the averaged similarity values. To do this, users were sampled using a Monte Carlo simulation and average similarity was calculated at the different levels. For example, to use the ASW method with two users, for each word pair average similarity values were calculated by selecting two random users and computing the average similarity. This is repeated for each word pair. Then, to compare to a single individual, the word pairs for that individual were attained and the correlation was taken to the ASW method with two users. This

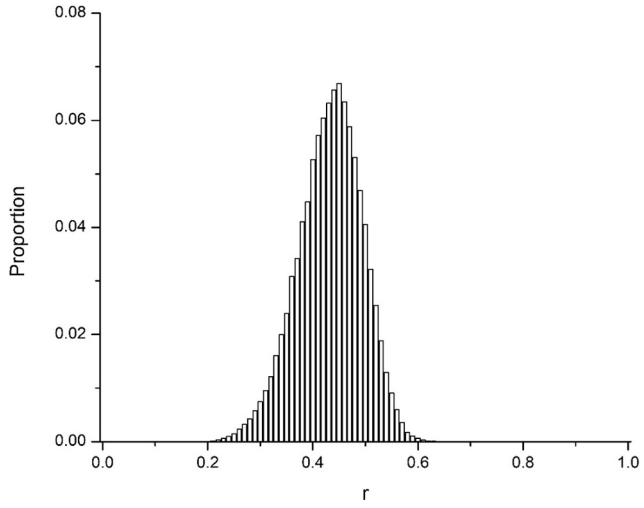

Fig. 3. A histogram of the correlations between the word pair similarity values for each user pair. The histogram consists of 124,750 correlations by comparing all 500 users to each other.

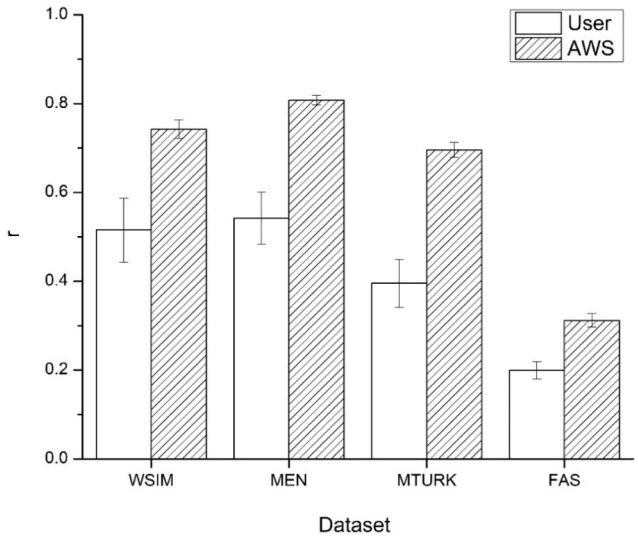

Fig. 4. A comparison of the fit between the individual user corpora to the fit of the ASW aggregation method. All 500 users were tested, and the ASW method was evaluated on the same word pairs that an individual was evaluated upon (i.e., the word pairs that exceeded the frequency parameter). Error bars are standard deviation.

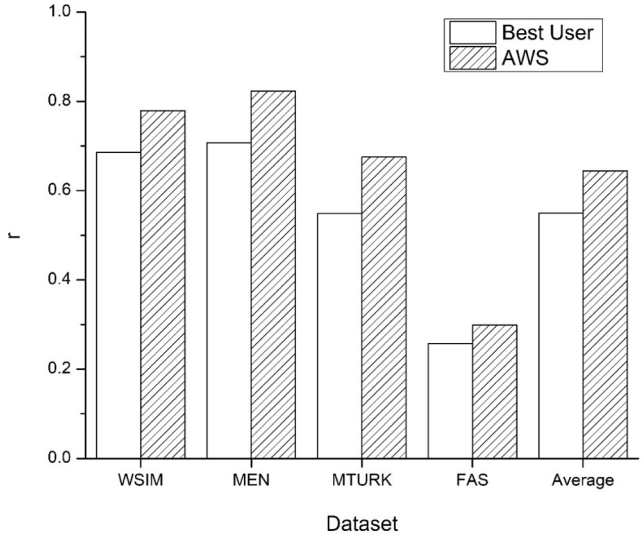

Fig. 5. A comparison of the best fitting individual for each word similarity dataset with the corresponding fit of the ASW method to that individual's word pairs.

was done 500 times with different users being randomly selected each time, and the average correlation for that individual's word pairs was computed (that is, there will be 500 correlations for each level of user inclusion). This was done for all 500 users. This process was repeated for all five levels of user inclusion, which will allow for the impact of an increasing number of users being aggregated has on the ASW method performance to be determined.

The result of this simulation is contained in Fig. 6. This figure shows that there is a dramatic increase in the fit of the ASW method as more users are included in the average similarity values, even with only two users included. This suggests that incorporating even a small number of users to calculate average similarity values leads to substantial improvement over the individual user representations. There is continued improvement across number of users samples but reaches asymptote at about 8–10 users, suggesting that the ASW method does not require all user to exceed the performance of individual users, but even a small number of individual word similarity ratings will suffice.

The correlations contained in Fig. 6 were calculated using only the word pairs for each individual user, which means that not all word pairs are being evaluated for the model. Thus, an additional simulation was done where the improvement of model fit was calculated for all word pairs where at least 30 users had produced similarity values, in order to determine the shape of improvement with a greater number of users. This was done by attaining all the word pairs for each dataset that had at least 30 individual similarity values. Then average similarity values for all word pairs were computed using a Monte Carlo simulation where the number of users who contributed to the average similarity values was randomly sampled without replacement. Specifically, average similarity values from 1 to 30 users, in step of 1, for all word pairs were calculated, with the users who contributed to the similarity values being randomly sampled. The fit of these similarity values to the datasets was then calculated. As in the previous simulation, this was repeated 500 times to calculate the average correlation at each sampled level.

The result of this simulation is contained in Fig. 7 for each dataset and shows, similar to Fig. 6, that there is considerable improvement in fit from 1 to 10 users, at which point the improvement plateaus and small refinements are made to fit. However, many of the word pairs used in the Fig. 7 contain significantly more than 30 user similarity values.

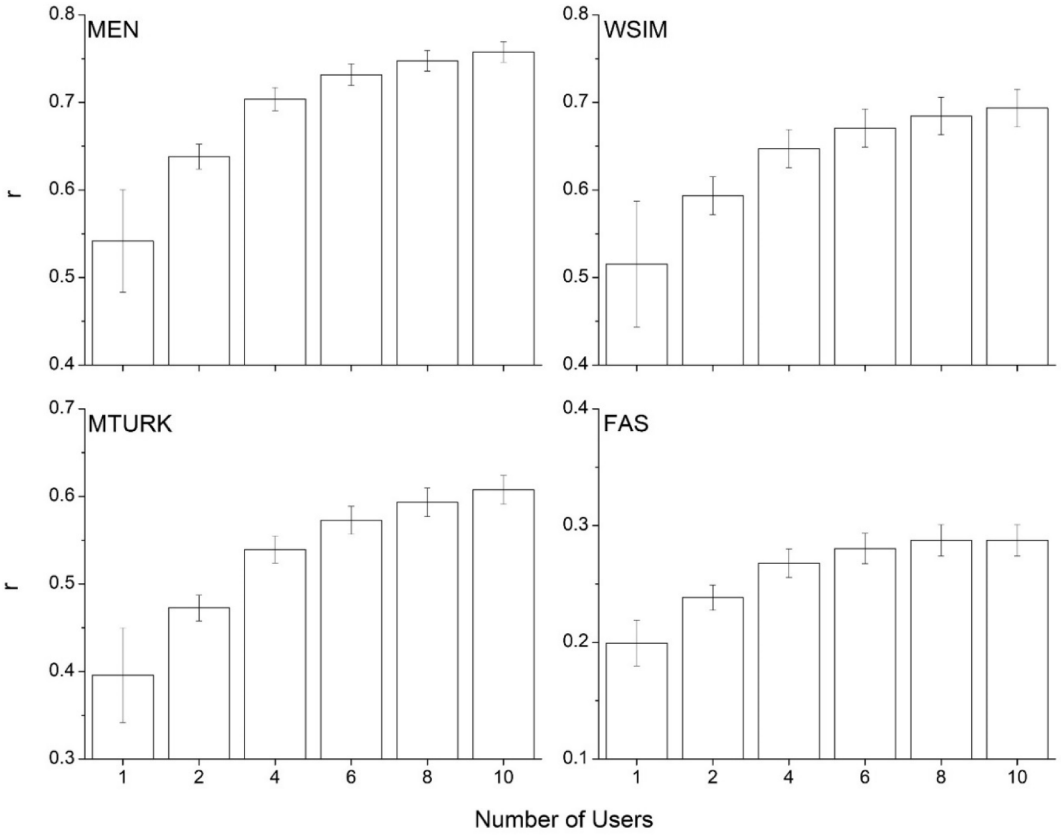

Fig. 6. The fit of the ASW method to each individual's word pair set as a function of the number of word similarity ratings included in the aggregation method. The ASW values were determined through a Monte Carlo simulation where the specified number of users were randomly selected across 500 samples. Error bars are standard deviation.

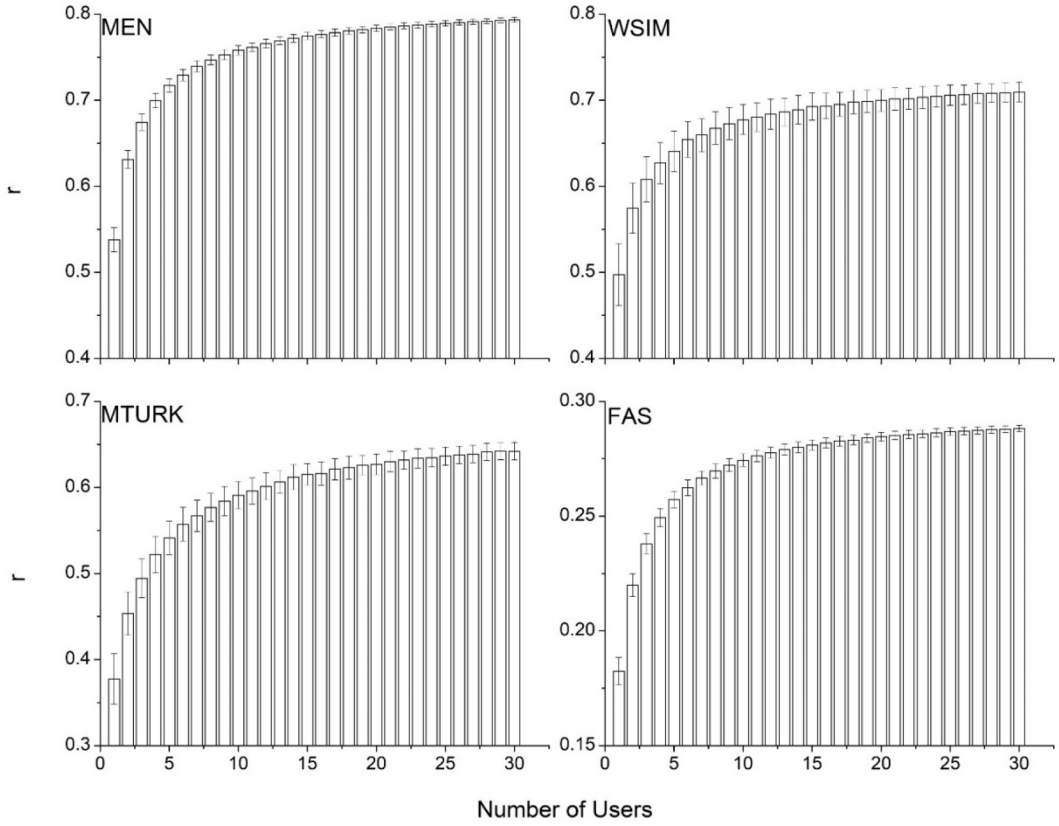

Fig. 7. The fit of the ASW method to all word pairs that had at least 30 users who produced similarity values for that pair as a function of the number of word similarity ratings included in the aggregation method. The ASW values were determined through a Monte Carlo simulation where the specified number of users were randomly selected across 500 samples. Error bars are standard deviation.

For example, for the MEN dataset each word pair contained on average 158.5 user similarity values. To determine to what extent the small refinement offered by additional users to the fit of the model, Fig. 8 contrasts the correlations between the average values for the 30 sampled users from Fig. 7 and the correlation when the average similarity is calculated for all users. This figure shows that there is a consistent increase in fit when all users are contained in the analysis, suggesting that maximizing the number of users provides a more precise estimation of word similarity.

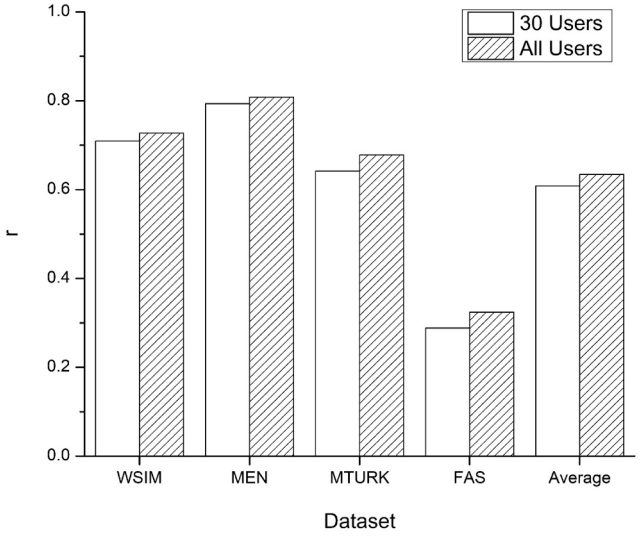

Fig. 8. A comparison of the fit of the ASW method when only 30 users are used in the aggregation method versus all users being included.

The final simulation performed will determine the impact that the uniqueness of each individual user's word similarity values has on model performance by randomizing each user corpus. This simulation is inspired by recent work by Hollis (2020) and Johns and Jones (2022) who examined the impact of randomized corpora when calculating contextual diversity measures of lexical strength. Similar to the research described here, Johns and Jones (2022) used a distributional model-based measure of contextual diversity (first described in Johns, 2021a) that was built through the analysis of individual user corpora to examine lexical organization data. Importantly, it was found that the measure did not provide a large advantage over other, more traditional measures (such as word frequency), when the uniqueness of each individual user's corpus was negated through randomization of the user corpora. This finding suggests that the unique pattern of language usage that each individual user has is important in accounting for lexical organization. Whether this same advantage applies to lexical semantics will be evaluated here.

Following Johns and Jones (2022), here all sentences from each individual corpus were collated into a single set. These sentences were then randomly split into 500 individual corpora of equal size. Thus, the intact and randomized corpora have the equivalent amount of linguistic information in them, just organized differently. The use of the randomized corpora entails that the corpora will not contain the unique signal of word meaning that each individual user provides. Instead, each of the randomized corpora contain a snapshot of all the language used across users, meaning the uniqueness of representations will be significantly reduced. The ASW method was used to calculate word similarity for each corpus set. Only word pairs that had at least 30 individual similarity values from each corpus set were included in the analysis. The result of this simulation is contained in Fig. 9, which shows that the ASW

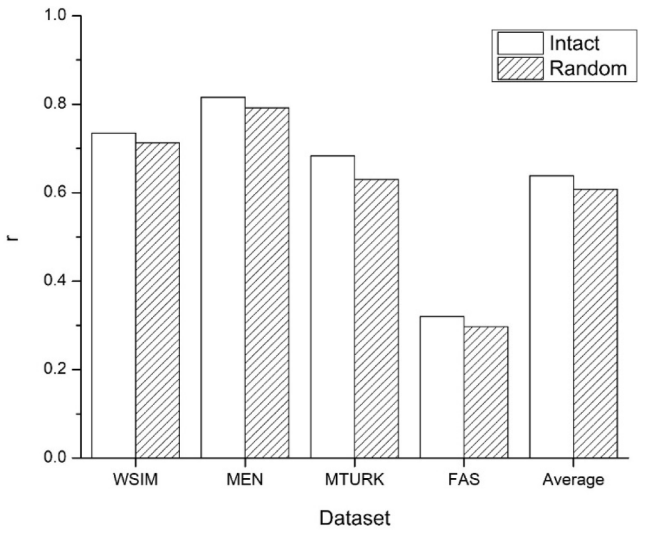

Fig. 9. A comparison of the ASW method for all users when the corpora are intact or randomized.

similarity values from the intact corpora outperforms the randomized corpora across every dataset. Although a relatively small effect, especially compared to the results of Johns and Jones (2022) in lexical organization, this finding suggests that the uniqueness of each individual user's usage of language, and the representation of word meaning derived from it, provides more accurate estimates of word meanings than those obtained from the randomized corpora which lack unique signals of individual differences in their lexical representations.

4. General discussion

The goal of this article was to determine if there is a wisdom of the crowd effect in lexical semantic memory, such that the aggregated word similarity values from many users exceeds the fit of the best fitting individual across multiple sets of word similarity and free association data. This was accomplished by training 500 different distributional models from 500 high-level commenters on the internet forum Reddit. It was found that averaging the word similarity values across the individual user models (the ASW method) provided superior performance than training a single model with all combined user corpora (the TR method). Using the ASW method, a strong wisdom of the crowd effect was found such that the aggregated similarity values far exceeded the average fit of the individual user models and also exceeded the performance of the best fitting individual for each dataset tested. Additionally, it was found that aggregating only a small number of users provided a large increase in fit relative to the individual corpora, but with the best fitting measure including word similarity values from all possible users. Finally, when the user corpora were randomized, a drop in model performance was found, suggesting that the uniqueness of word meanings across individuals plays an important role in determining accurate estimations of word meanings.

Linguistic experience varies across individuals and is impacted by both demography and personal preferences. Variability in linguistic experience results in word meanings being different across cultures (Thompson et al., 2020) and people (Johns, 2022). As Fig. 2 demonstrates, the word meanings that were derived from the individual corpora used here vary appreciatively from each other, suggesting that each individual user corpus encodes different forms of a word's meaning. A distributional model derived from only one user corpus was found here to provide relatively poor performance, but once the word similarities from all user corpora were aggregated together a good accounting of word similarity data was found. From a theoretical perspective, this finding suggests that a target for future research in lexical semantics should be focused on attaining a better understanding of the underlying variability of word meanings across a population of language users and how this knowledge can be used to generate better, and more cognitively plausible, models of lexical semantics.

The distributional modeling approach of using multiple (and small) individual corpora to evaluate the wisdom of the crowd effect in lexical semantics runs counter to the standard training methodology used for this model type. Typically, when training a distributional model, the amount of training materials that is given to a model is maximized to ensure that the model has acquired as much linguistic knowledge as possible. Here, the similarity values from many small corpora, attained from individual language users, were aggregated to form average word similarity values, and these aggregated similarity values outperformed a single model trained on the combined individual corpora. This suggests that the unique representations that each individual language user has provides unique information about the meaning of a word. These unique perspectives are reduced in importance when all language is combined (or the individual corpora are randomized, as is shown in Fig. 9). Thus, one potential pathway towards producing better distributional models is not to generate models that can learn and represent huge amounts of linguistic materials, but instead models that cohere more to individual levels of knowledge. Future model development on this path will require a rethinking of both the types of learning mechanisms that are at play in distributional learning (given that models are impacted significantly by corpus size; Levy, Goldberg, & Dagan, 2015) and also the training materials that are used, with the individual Reddit corpora used here providing a promising starting point for the development of individual-level distributional models.

In the ASW method used in this article, the word similarity values were simply averaged across users. One method that has proven effective at improving the performance of the crowd is to reduce or remove the poorest performing members (Bennett et al., 2018; Budescu & Chen, 2015), with methods that incorporate the relative expertise of individuals outperforming standard aggregation techniques (Lee, Steyvers, & Miller, 2014). When examining lexical semantic data, it is difficult to determine who "expert" users of language are, such that an aggregation algorithm could preferentially choose the best user corpus given that the model is fitting to simple psycholinguistic behavioral data. However, previous research in distributional modeling has demonstrated that by curating the text that is used to train a model can significantly improve model performance. For example, Johns, Jones, and Mewhort (2019; see also Johns & Jamieson, 2019 for an example of its use) developed an optimization framework for distributional models that used supervised learning to choose the selection of texts that maximizes model performance, which resulted in large performance increases for experience-based cognitive models across multiple behavioral datatypes. This same procedure could be used to determine the best set of users that provides the best fit to a set of data rather than use all possible ratings to optimize the word similarity values that are produced with an aggregation method.

The results of this article, as well as previous results (e.g., Johns, 2021a,b, 2022; Johns & Jones, 2022) speak to the advantages of using Reddit, and social media more generally (see Herdağdelen and Marelli, 2017; Johns, 2019; Otto, Devine, Schulz, Bornstein, & Louie, 2022; Otto & Eichstaedt, 2018 for other examples), as a source of training materials for distributional cognitive models. Social media is becoming an increasingly popular means of communication, especially among younger people, where at least 92% of teenagers have been found to be active on a social media site (Lenhart, 2015). Given that people use social media as a form of communication with their social group, this suggests that corpora derived from these sources are more naturalistic in terms of the language contained in them, at least compared to other standard corpus types used such as Wikipedia, newspaper articles, or books. Additionally, as explored here, the metadata attached to comments allows for insight into the individual variability inherent in the usage of language to be measured. Given the widespread success of

distributional modeling to theoretical issues in the cognitive and psychological sciences using standard methodologies (Bhatia & Aka, 2022; Günther, Rinaldi, & Marelli, 2019; Kumar, 2020), a goal in this field of research should be to bridge the connection between the overall language environment (which is what models trained on very large corpora are representing) down to the level of the individuals who produced that language environment and learned from it. Analysis of social media sources provides a promising avenue to pursue this question.

Although lexical semantics was the only data type examined here, the application of using multiple individual corpora to drive cognitive models could have widespread utility. For example, Bhatia and colleagues (e.g., Bhatia, 2017, 2019; Bhatia & Walasek, 2019; Zou & Bhatia, 2021; for a review, see Bhatia, Richie, & Zou, 2019) have demonstrated that distributional models are capable of accounting for widespread data in judgement and decision making. Given the significant levels of individual differences inherent in this literature (e.g., Stanovich, 1999), by taking multiple distributional models trained from different individual language users, a better understanding of the impact of the structure language environment has on judgement across people could be attained. An additional area where distributional models have been used is in the cognitive modeling of episodic memory (e.g., Johns, Jones, & Mewhort, 2012, 2021; Osth, Shabahang, Mewhort, & Heathcote, 2020; Mewhort, Shabahang, & Franklin, 2018; Reid & Jamieson, 2023), where the representations derived from a distributional model are used in conjunction with a process model of episodic memory to account for the impact of lexical semantics on memory performance. However, there is significant variability in item-level performance in memory, both in true (e.g., Cortese, Khanna, & Hacker, 2010; Cortese, McCarty, & Schock, 2015) and false (e.g., Gallo & Roediger, 2002; Stadler, Roediger, & McDermott, 1999) memory. Examining individual-level differences in the representations of words that are used in memory experiments could potentially explain this variance in behavior.

The results of this article are also related to recent theoretical discussions centering around the nature of the relativity of word meanings across and within languages, following the work of Thompson et al. (2020) and Johns (2022). Using distributional modeling techniques, Thompson et al. (2020) found that different languages produce significantly different representations for the same word, while Johns (2022) found that there was significant variability in the meaning of words in the same language at the individual user level, although not to the same extent that was found in the across language examination of Thompson et al. (2020). These previous results suggest that word meanings are malleable both at the cultural and individual level. The results of this article point to the notion that to understand the average meaning of a word that users of a language have requires modeling techniques that take into account the underlying variability of word meanings that individuals have. Every individual does not have the same underlying representation of a word's meaning and so to produce accurate computational models of lexical semantics requires taking this variability into account when calculating word similarities, which current approaches in distributional modeling fail to do due to the use of very large corpora that lack ecological validity. Big data methodologies have revolutionized the study of lexical semantics (Johns, Jamieson, & Jones, 2020) but to truly understand the nature of the underlying cognitive processes involved in knowledge acquisition, representation, and use requires moving down to a level of analysis to the individual.

A key requirement for the validation of this technique will be to use targeted experimentation to determine the power that using individualized models allows for in accounting for behavioral data in comparison to more standard distributional techniques. Ideally, one would be able to recruit high-level commenters on social media, attain a corpus of their comments, evaluate their performance on a variety of psycholinguistic tasks (e.g., word similarity, free association, and lexical decision tasks), and determine how well their own corpus fits to their resulting behavior as compared to another user's corpus. However, this is likely infeasible due to a number of obvious ethical and logistical challenges. A more realistic possibility is to recruit participants who consume and produce comments on social media (e.g., Reddit) at a high rate and determine the sections of the site where they communicate (e.g., the specific subreddits that they interact on). Then, the corpora that best map onto their commenting patterns could be identified and the fit of those corpora would be evaluated in contrast to user corpora that have a different commenting pattern. This would provide a quantitative examination into the advantage that accounting for individual-level linguistic experience provides in accounting for the variability of lexical behavior, a fundamental question that machine learning approaches to cognitive modeling are uniquely capable of answering (Johns, Jamieson, & Jones, in press).

However, there are many challenges in developing this individualized approach to distributional modeling. From a technical perspective, the approach developed here is computationally burdensome as compared to standard techniques, as it requires training, storing, and integrating similarity values across a large number of different models. These issues are somewhat alleviated by the use of a computer with a large amount of main memory in order to maintain multiple models concurrently (e.g., the computer that was used for the simulations in this article had 256gb of RAM), which may limit the adoption of these techniques. However, a more pertinent issue is one of data reduction, as this article only used models from 500 individual users (for comparison's sake, the models described in Johns, 2021a utilized data from over 330,000 different users). This was a necessary sacrifice, as the techniques developed by Johns et al. (2019), and neural embedding models more generally, require relatively large samples of language to form stable word representations, mainly due to requiring high baseline word frequency levels to estimate the uniqueness of co-occurrence trends across words in a corpus. Thus, each user corpus had to be of sufficient size to be utilized by the methods described here. However, there are other techniques that are likely more resilient to lower word frequency values, such as pointwise mutual information (Bullinaria & Levy, 2007, 2012), an information theoretic measure of the probability of two words occurring in context together. Although this metric tends to perform worse than neural embedding models (Mandera et al., 2017), it is possible that computing the PMI of word pairs across a very large number of user corpora could provide better word similarity estimates than computing PMI from a larger, non-individualized corpus. However, whether this approach would provide a better fit to utilizing fewer corpora with more advanced distributional models is an important topic for future research.

Distributional models of semantics have demonstrated the systematic connection between the language that people experience and lexical behavior. Standard methodological practice in distributional model training calls for the use of very large corpora to generate representations of word meanings. However, this practice ignores the unique representations of word meanings that each individual user of a language has. The results of this article demonstrate that aggregating word similarity values across individual user models results in superior performance compared to any single individual user corpus, producing a pronounced wisdom of the crowd effect in lexical semantics. This finding suggests that a promising avenue for future distributional model development is to generate models that capture knowledge at the individual level and new aggregation techniques that can determine word meanings at the group level. However, for this theoretical pathway to succeed more emphasis needs to be placed on understanding individual variance in the usage and comprehension of language, an important goal for the cognitive and language sciences.

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.

Data availability

No data was used for the research described in the article.

Acknowledgements

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

References

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 417–439). Norwell, MA: Kluwer Academic.

Aujla, H. (2021). Language experience predicts semantic priming of lexical decision. *Canadian Journal of Experimental Psychology*, 75, 235-244.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review, 124*, 1–20.

Bhatia, S. (2019). Semantic processes in preferential decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*, 627–640.

Bhatia, S., & Walasek, L. (2019). Association and response accuracy in the wild. *Memory & Cognition, 47*, 292–298.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences, 29*, 31–36.

Bartlett, F. C. (1928). An experiment upon repeated reproduction. *Journal of General Psychology, 1*, 54–63.

Bartlett, F. C. (1932). *Remembering: An experimental and social study*. Cambridge: Cambridge University.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 14, pp. 830–839).

Bennett, S. T., Benjamin, A. S., Mistry, P. K., & Steyvers, M. (2018). Making a wiser crowd: Benefits of individual metacognitive control on crowd performance. *Computational Brain & Behavior, 1*, 90–99.

Bergman, E. T., & Roediger, H. L. (1999). Can Bartlett's repeated reproduction experiments be replicated? *Memory & Cognition, 27*, 937–947.

Bhatia, S., & Aka, A. (2022). Cognitive modeling with representations from large-scale digital data. *Current Directions in Psychological Science, 31*, 207–214.

Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012). In *Distributional semantics in technicolor* (pp. 136–145). Association for Computational Linguistics.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39*, 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*, 890–907.

Cortese, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory, 18*, 595–609.

Cortese, M. J., McCarty, D. P., & Schock, J. (2015). A mega recognition memory study of 2897 disyllabic words. *Quarterly Journal of Experimental Psychology, 68*, 1489–1501.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*, 159–190.

De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. In *COLING* (pp. 1861–1870). Association for Computational Linguistics: Stroudsburg, PA.

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.

Gallo, D. A., & Roediger, H. L., III (2002). Variability among word lists in eliciting memory illusions: Evidence for associative activation and monitoring. *Journal of Memory and Language, 47*, 469–497.

Galton, F. (1907). *Vox Populi. Nature, 75*, 450–451.

Gordon, K. (1924). Group Judgments in the Field of Lifted Weights. *Journal of Experimental Psychology, 7*, 398–400.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*, 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*, 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*, 976–995.

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*, Article 104146.

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior, 1*, 119–136.

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. (2021a). Disentangling contextual diversity: Communicative need as a lexical organizer. *Psychological Review, 128*, 525–557.

Johns, B. T. (2021b). Distributional social semantics: Inferring word meanings from communication patterns. *Cognitive Psychology, 131*, 10144.

Johns, B. T. (2022). *Determining the relativity of word meanings through the construction of individualized models of semantic memory*. PsyArXiv.

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*, 841–855.

Johns, B. T., & Jamieson, R. K. (2018). A large-scale analysis of variance in written language. *Cognitive Science, 42*, 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., 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., & Jones, M. N. (2022). Content matters: Measures of contextual diversity must consider semantic content. *Journal of Memory and Language, 123*, Article 104313.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology, 65*, 486–518.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review, 26*, 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., Mewhort, D. J. K., & Jones, M. N. (2019). The role of negative information in distributional semantic learning. *Cognitive Science, 43*, e1273.

Johns, B. T., Jamieson, R. K., & Jones, M. N. (in press). Scalable cognitive modeling: Putting Simon's (1969) ant back on the beach. *Canadian Journal of Experimental Psychology*.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*, 1–37.

Kumar, A. A. (2020). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, 1–41.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–224.

Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making, 9*, 259–273.

Lee, M. D., Steyvers, M., & Miller, B. (2014). A cognitive model for aggregating people's rankings. *PloS One, 9*, e96431.

Lee, M. D., Zhang, S., & Shi, J. (2011). The wisdom of the crowd playing The Price Is Right. *Memory & Cognition, 39*, 914–923.

Lenhart, A. (2015). *Teens, social media & Technology*. Pew Research Center: Retrieved from. http://www.pewinternet.org/2015/04/09/teens-social-mediatechnology-2015.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. 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.

Maki, W. S. (2008). A database of associative strengths from the strength-sampling model: A theory-based supplement to the Nelson, McEvoy, and Schreiber word association norms. *Behavior Research Methods, 40*, 232–235.

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

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*, 276–299.

Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2017). A neglected dimension of good forecasting judgment: The questions we choose also matter. *International Journal of Forecasting, 33*, 817–832.

Mewhort, D. J. K., Shabahang, K. D., & Franklin, D. R. J. (2018). Release from PI: An analysis and a model. *Psychonomic Bulletin & Review, 25*, 932–995.

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

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

Osth, A. F., Shabahang, K. D., Mewhort, D. J., & 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*, Article 104071.

Otto, A. R., Devine, S., Schulz, E., Bornstein, A. M., & Louie, K. (2022). Context-dependent choice and evaluation in real-world consumer behavior. *bioRxiv*.

Otto, A. R., & Eichstaedt, J. C. (2018). Real-world unexpected outcomes predict city-level mood states and risk-taking behavior. *PloS One, 13*, e0206923.

Reid, J. N., & Jamieson, R. K. (2023). True and false recognition in MINERVA 2: Extension to sentences and metaphors. *Journal of Memory and Language, 129*, Article 104397.

Shabahang, K. D., Yim, H., & Dennis, S. J. (2022). Generalization at retrieval using associative networks with transient weight changes. *Computational Brain & Behavior, 5*, 124–155.
Shaoul, C., & Westbury, C. (2010). *The Westbury Lab Wikipedia Corpus 201*. Edmonton, AB: University of Alberta.
Singh, M., Richie, R., & Bhatia, S. (2022). Representing and predicting everyday behavior. *Computational Brain & Behavior, 5*, 1–21.
Stadler, M. A., Roediger, H. L., & McDermott, K. B. (1999). Norms for word lists that create false memories. *Memory & Cognition, 27*, 494–500.
Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning*. Psychology Press.
Steyvers, M., Lee, M. D., Miller, B., & Hemmer, P. (2009). The wisdom of crowds in the recollection of orderinformation. In J. Lafferty, & C. Williams (Eds.), *Advances in neural information processing systems, 23* (pp. 1785–1793). Cambridge, MA: MIT Press.
Steyvers, M., & Miller, B. (2015). Cognition and collective intelligence. In T. W. Malone, & M. S. Bernstein (Eds.), *Handbook of collective intelligence* (pp. 119–137). MIT Press.
Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*. New York: Random House.
Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour, 4*, 1029–1038.
Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science, 13*, 75–78.
Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science, 36*, 452–470.
Zou, W., & Bhatia, S. (2021). Judgment errors in naturalistic numerical estimation. *Cognition, 211*, Article 104647.