Context Matters: Recovering Human Semantic Structure from Machine Learning Analysis of Large-Scale Text Corpora

Marius Cătălin Iordan, Tyler Giallanza, Cameron T. Ellis, Nicole Beckage, Jonathan D. Cohen*

Contact Information

Marius Cătălin Iordan, Ph.D.
Princeton Neuroscience Institute & Psychology Department, Princeton University
Email: mci@princeton.edu
**corresponding author*

Tyler Giallanza
Computer Science Department, Southern Methodist University
Email: tgiallanza@smu.edu

Cameron T. Ellis
Psychology Department, Yale University
Email: cameron.ellis@yale.edu

Nicole Beckage, Ph.D.
Intel Research
Email: nicole.beckage@intel.com

Jonathan D. Cohen, MD, Ph.D.
Princeton Neuroscience Institute & Psychology Department, Princeton University
Email: jdc@princeton.edu

Manuscript Information

Number of Words	
Title	14 / 15
Abstract	150 / 150
Main Text	4,887 / 5,000
Methods	2,583 / 3,000
Number of References	55 / 70
Number of Display Items	5 / 10

Abstract

Understanding how human semantic knowledge is organized and how people use it to judge fundamental relationships, such as similarity between concepts, has proven difficult. Theoretical models have consistently failed to provide accurate predictions of human judgments, as has the application of machine learning algorithms to large-scale, text-based corpora (embedding spaces). Based on the hypothesis that context plays a critical role in human cognition, we show that generating embedding spaces using contextually-constrained text corpora greatly improves their ability to predict human judgments. Additionally, we introduce a novel context-based method for extracting interpretable feature information (e.g., size) from embedding spaces. Our findings suggest that contextually-constraining large-scale text corpora, coupled with applying state-of-the-art machine learning algorithms, may improve the correspondence between representations derived using such methods and those underlying human semantic structure. This promises to provide novel insight into human similarity judgments and designing algorithms that can interact effectively with human semantic knowledge.

Introduction

Understanding the underlying structure of human semantic representations is a fundamental and longstanding goal of cognitive science[1,2,3,4,5,6,7], with implications that range broadly from neuroscience[8,9] to computer science[10,11,12,13] and beyond[14]. Most theories of semantic knowledge propose that items in semantic memory are represented in a multidimensional feature space, and that key relationships among items – such as similarity and category structure – are determined by distance among items in this space[1,2,4,15,16,17,18,19]. However, defining the space, establishing how distances are determined, and using these distances to predict human judgments about semantic relationships remains a challenge. The best efforts to define theoretical principles (e.g., formal metrics) that can predict semantic judgments from feature representations[3,16,20,21,22,23,24] capture less than half the variance observed in empirical studies of such judgments. At the same time, a comprehensive empirical determination of the structure of human semantic representation (e.g., by evaluating all possible similarity relationships) is impossible, given that human semantic knowledge encompasses tens of thousands of categories and trillions of individual objects[25].

Complementing research in cognitive psychology, machine learning – and work on natural language processing (NLP) in particular – has attempted to use large amounts of human generated text to create a high dimensional representation that

may provide insights into the semantic space humans use when processing and producing language (i.e., millions or billions of words[13,26,27,28]). These approaches generate multidimensional vector spaces learned from the statistics of the input articles, in which words that appear in similar syntactic contexts across different sources of writing (e.g., sentences, articles, books) are placed close to one another, and words that occur in less similar context are places farther apart. A distance metric between a given pair of words can then be used as a measure of their similarity. This approach has met with some success in predicting categorical distinctions[29], predicting properties of objects[30,31,32], and even revealing cultural stereotypes and implicit associations hidden within the documents[14]. However, the measures of similarity generated by such machine learning methods have remained limited in their ability to predict direct empirical measurements of human similarity judgments[10,30].

Thus, neither the top-down theoretically-principled approaches, nor bottom-up data-driven approaches have provided a full, empirically validated understanding of the structure of human semantic representations. Here, we consider a novel approach that uses machine learning models to predict human empirical judgments, motivated by the hypothesis that the latter are heavily influenced by the context in which they are made[33,34,35,36,37,38]. This attentional influence can include task demands (e.g., instructions provided by experimenters), incidental factors related to the circumstances of the task, and/or features of the items to be judged. For example, when asked to judge the similarity between a bear and a bull among a number of other animals, attention may be directed to their physical characteristics as objects in a natural context (e.g., size, domesticity), leading to the judgment that they are similar; however, in the context of financial markets, attention may be drawn to their symbolic meaning or economic value, leading to the judgment that they are very different.

Motivated by the idea that attention has a strong influence on semantic judgments, we modified the use of modern machine learning methods for generating data-driven high-dimensional semantic embedding spaces by introducing semantic contextual constraints in the construction of the text corpora from which the embedding spaces are learned. These constraints are intended to parallel the contextual constraints that might impact human similarity judgments by restricting the corpora to materials generated by individuals (e.g., authors or speakers) whose mental context aligns more closely with that of the participants whose judgments are measured in empirical studies. We test this approach in three experiments.

The first two experiments demonstrate that embedding spaces learned from contextually-constrained text corpora substantially improve the ability to predict empirical measures of human semantic judgments (pairwise similarity judgments in Experiment 1 and item-specific feature ratings in Experiment 2). In the third experiment, we test the usefulness of contextually-relevant features for generating

distance metrics in context-free embedding spaces that can better recover empirical similarity judgments from such spaces, though they still cannot match the performance of contextually-constrained corpora. Finally, we show that applying a similar method to embeddings from contextually-constrained corpora provides the best prediction of human similarity judgments achieved to date, exceeding 90% of human inter-rater reliability in two specific semantic contexts.

Collectively, our results support the hypothesis that human semantic processing is fundamentally constrained by contextual influences and that attempts to study and understand semantic knowledge (whether in laboratory experiments or by using large-scale text corpora) must take account of these attentional influences. By leveraging the application of machine learning methods to large-scale text corpora, our work provides an approach to studying complex cognitive processes at a scale that may be difficult to achieve with laboratory studies. This may help accelerate both our understanding of the underlying structure of human semantic representations, as well as efforts to build artificial systems that can emulate and/or better interact with this structure.

Results

Experiment 1a: Contextually-Constrained Semantic Embeddings Capture Human Similarity Judgments Better than Context-Free Embeddings

Word embedding spaces are generated by training machine learning models on large corpora of text, often using deep neural network algorithms. This approach is typically applied to the largest corpora available, on the assumption that larger datasets will provide more accurate estimation of the underlying semantic structure. However, aggregating across multiple contexts (e.g., National Geographic and Wall Street Journal) may dilute the sensitivity of resulting embedding spaces to attentionally-constrained human semantic judgments. To test whether contextually-constraining the corpora used to train machine learning algorithms improves their ability to predict empirical similarity judgments, we collected Wikipedia articles related to two distinct semantic contexts (Fig. 1a): 'nature' (~70 million words) and 'transportation' (~50 million words). We then generated contextually-constrained embedding spaces by training continuous skip-gram Word2Vec models with negative sampling[26,27] using the two collections of contextually-relevant articles as training sets. We chose Word2Vec to train our embedding spaces because this type of model has been shown to be superior to other embedding models in matching human cognitive judgments[30].

We compared the two contextually-constrained embedding spaces to a context-free Word2Vec embedding space trained on all English language Wikipedia articles (~2 billion words) and to an embedding space trained on a random subset of this training corpus, size-matched to the contextually-constrained embedding spaces (~60 million words). We also compared performance of the four Word2Vec embedding spaces to another commonly used embedding space known as GloVe[28] for two main reasons; first, the GloVe embeddings are learned from the Common Crawl corpus (~42 billion words) and thus provide insight into the role of corpus size on making predictions about human judgments; and, second, GloVe is a widely used independent algorithm for learning embeddings that allows us to evaluate the generality of our results beyond Word2vec. Finally, we compared our performance against a recent object similarity model (SPoSE[39]), which is the largest scale attempt, to date, of using machine learning models to predict relationships between semantic concepts.

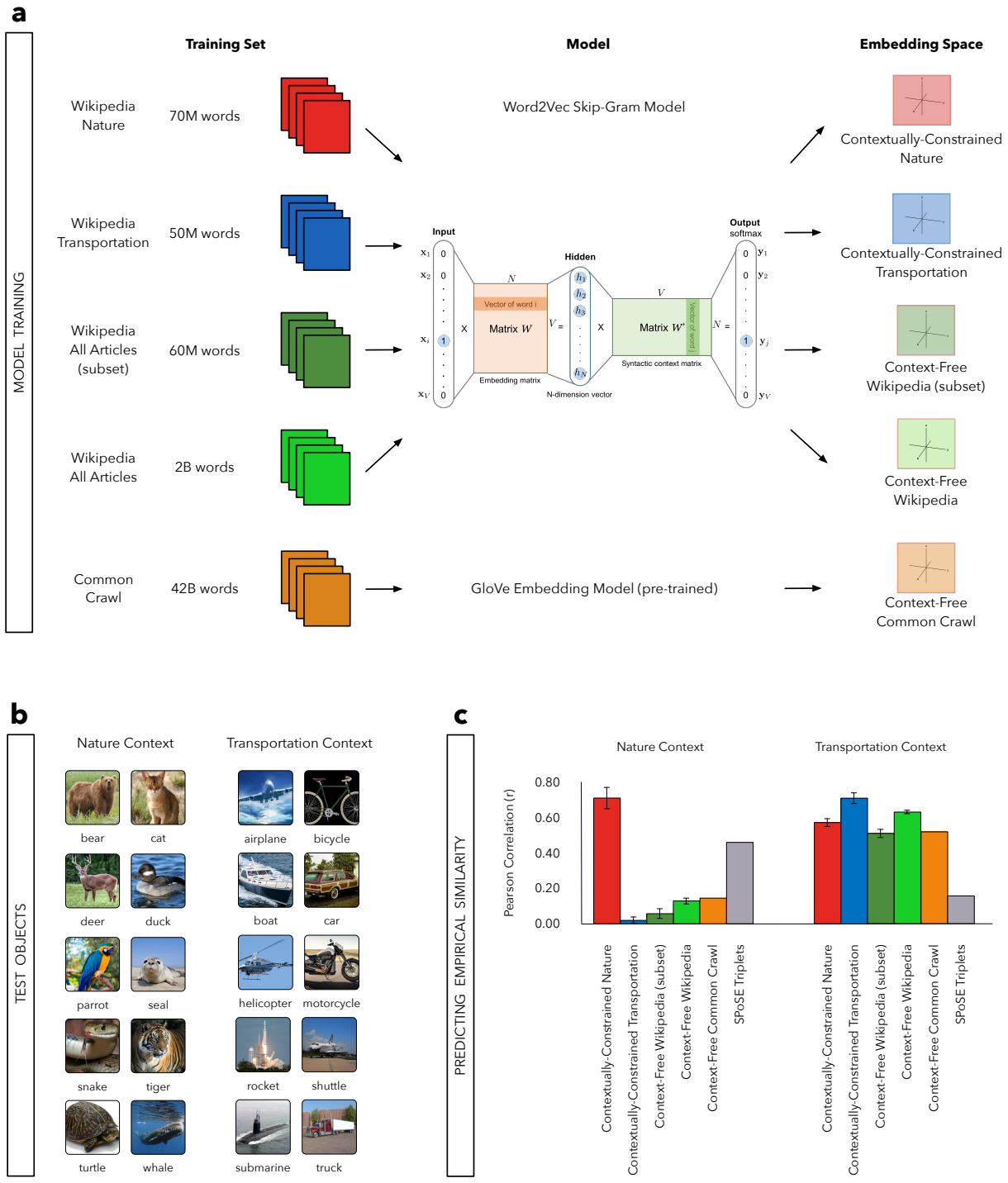

Figure 1. Generating Contextually-Constrained Embedding Spaces and Testing Their Alignment with Human Similarity Judgments. *(a) Model Training. We generated contextually-constrained embedding spaces using training sets comprised of Wikipedia articles considered relevant to each semantic context ('nature' ~70M words, 'transportation' ~ 50M words). Similarly, we trained context-free models with the*

training set of all publicly available Wikipedia articles (~2B words), as well as a size-matched subset of this corpus (~60M words). Finally, we compared results using these models with a pre-trained embedding space trained on the Common Crawl corpus (GloVe, ~42B words) and against a recent object similarity model (SPoSE). (b) To quantify how well each embedding space aligned with human similarity judgments for the two semantic test contexts, we selected 10 representative basic-level objects representative of each test context (10 animals and 10 vehicles) and collected human-reported similarity judgments between for all pairs of objects in each context (45 pairs per context). (c) We computed Pearson correlation between human empirical similarity judgments (all 45 pairwise comparisons within each semantic context, averaged across participants) and similarity predicted by each embedding model (cosine distance between embedding vectors corresponding to each object in each model). Error bars show 95% confidence interval for 10 independent initializations of the models used to generate embeddings. All differences between bars are statistically significant, $p<0.001$.

To test how well each embedding space could predict human similarity judgments, we selected two representative subsets of ten concrete basic-level objects[40] commonly associated with each of the two contexts (10 animals, e.g., 'bear'; 10 vehicles, e.g., 'car'; Fig. 1b). We then used the Amazon Mechanical Turk online platform to collect empirical similarity judgments on a Likert scale (1–5) for all pairs of objects within each context (e.g., by asking participants to judge "How similar are a bear and a cat?"; or "How similar are a bicycle and a car?"). Each participant made judgments in a single semantic context (i.e., all between-animal comparisons or all between-vehicle comparisons, but not both) and the order of comparisons in each context was counterbalanced across participants. After removing subject responses with low inter-rater reliability, we computed the empirical ground truth similarity for each pair of objects in each context as the average of these judgments across the remaining participants. We used cosine distance between embedding vectors corresponding to the 10 animals and 10 vehicles to generate similarity predictions for all pairwise combinations of objects in each of the embedding spaces. Finally, we calculated the Pearson correlation between the cosine distance measures and the empirical similarity judgments to assess how well each embedding space can account for human judgments of pairwise similarity.

For animals, estimates of similarity using the contextually-constrained nature embedding space were highly correlated with human judgments ($r=0.711\pm0.071$(CI); Fig. 1c). In contrast, estimates from the contextually-constrained transportation embedding space, the context-free models, and the SPoSE triplet space could not recover the same pattern of human similarity judgments among animals

(transportation r=0.019±0.028(CI); Wikipedia subset r=0.058±0.038(CI); Wikipedia r=0.129±0.027(CI); Common Crawl r=0.145; SPoSE r=0.460; nature > transportation $t(9)=66.3$, $p<0.001$; nature > Wikipedia subset $t(9)=55.3$, $p<0.001$; nature > Wikipedia $t(9)=61.9$, $p<0.001$; nature > Common Crawl $t(9)=54.0$, $p<0.001$; nature > SPoSE $t(9)=23.9$, $p<0.001$). Conversely, for vehicles, similarity estimates from its corresponding contextually-constrained transportation embedding space were the most highly correlated with human judgments (r=0.711±0.040(CI)). While similarity estimates from the other embedding spaces and from the SPoSE space were also highly correlated with empirical judgments of similarity among vehicles (nature r=0.572±0.032(CI); Wikipedia subset r=0.510±0.033(CI); Wikipedia r=0.631±0.020(CI); Common Crawl r=0.519; SPoSE r=0.156), the effects were significantly weaker that for the contextually-constrained transportation embedding space (transportation > nature $t(9)=18.4$,$p<0.001$; transportation > Wikipedia subset $t(9)=29.1$,$p<0.001$; transportation > Wikipedia $t(9)=14.3$,$p<0.001$; transportation > Common Crawl $t(9)=32.2$,$p<0.001$; transportation > SPoSE $t(9)=93.0$,$p<0.001$).

These results strongly support the hypothesis that human similarity judgments are subject to contextual influences and that such judgments can be better predicted by incorporating contextual constraints into the training procedure used to build word embedding spaces. This was evidenced in two major ways: first, each contextually-constrained embedding space made more accurate predictions about judgments for its corresponding category of objects compared to the context-free models (e.g., using embeddings built from articles corresponding to the nature context to predict similarity judgments about animals); second, we observed a double-dissociation between the performance of the contextually-constrained models according to context: predictions of similarity judgments were most substantially improved by using contextually-constrained corpora specifically when the contextual constraint aligned with the category of objects being judged. Furthermore, this double dissociation was robust across multiple hyper-parameter choices for the Word2Vec model, such as number of words that are part of the syntactic context during training (i.e., window size), or the dimensionality of the learned embedding spaces (Supplementary Figs. 1 & 2). The double dissociation was also observed when bootstrapping the test set pairwise comparisons, indicating that the difference in performance between the models was not disproportionally driven by particular animals or vehicles chosen for the test set (Supplementary Figs. 3 & 4).

Experiment 1b: Combining Semantic Contexts Degrades Contextually-Relevant Information

We hypothesized that context-specific embedding spaces would be better aligned with human judgments because they were built from corpora that reflect the influences of semantic contexts in the minds of the empirical study participants as they made similarity judgments. To further test this idea, we evaluated the extent to which cross-contextual contamination induces a misalignment between distances in embedding spaces and human similarity judgments. We generated new combined-context embedding spaces using different proportions of the training data from each of the two semantic contexts (nature and transportation; Fig. 2a), both matching for the size of the contextually-constrained models' training set (60M words; canonical combined-context model), as well as using all available training data from the two semantic contexts (120M words; full context-combined model).

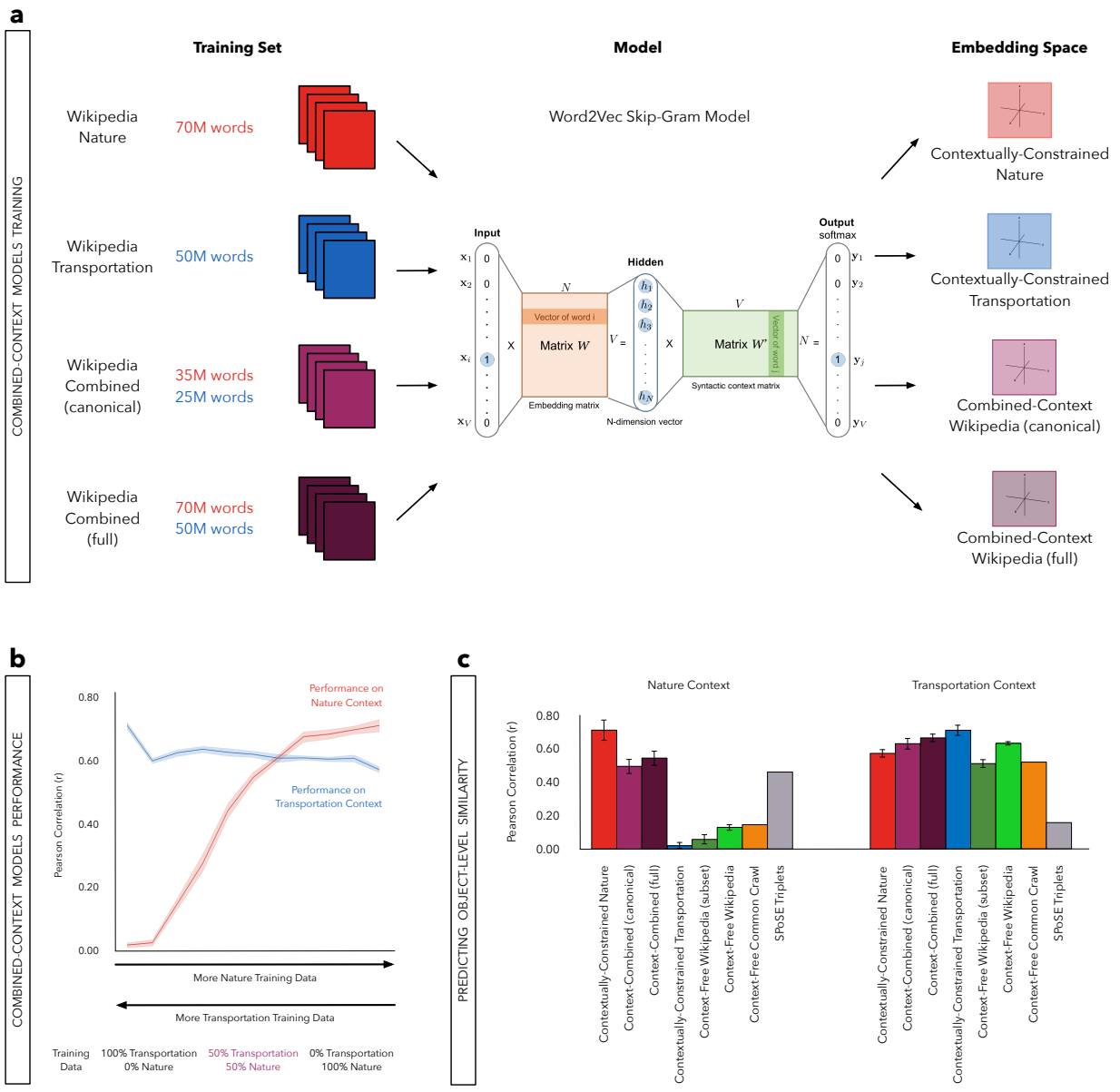

Figure 2. Context-Combined Models More Poorly Predict Similarity Judgments. *(a) Combined-context embedding spaces were generated by using training data from the nature and transportation semantic contexts in different proportions (60M words, e.g., 10%-90%, 50%-50%, etc.). A full context-combined embedding space was also generated using all available training data from both semantic contexts (120M words). (b) When combining training data from two semantic contexts at different ratios, the resulting combined-context embeddings recover a proportional amount of information from their preferred/non-preferred semantic contexts. (c) The canonical and full combined-context models produced distances between concepts that were less aligned with human judgments in both the nature and the transportation semantic*

contexts, respectively, compared to the corresponding contextually-constrained embedding spaces. Errors signify 95% confidence intervals for 10 independent initializations of the embedding models. All differences between contextually-constrained models in their preferred context and other models are statistically significant, p<0.001.

As predicted, we found that performance of the combined-context embedding spaces was intermediate between the preferred and non-preferred contextually-constrained embedding spaces in predicting human similarity judgments: as more nature semantic context data was used to train the combined-context models, the alignment between embedding spaces and human judgments for the animal test set improved; and, conversely, more transportation semantic context data yielded better recovery of similarity relationships in the vehicle test set (Fig. 2b). We illustrated this performance difference using the 50% nature-50% transportation embedding spaces in Fig. 2c (nature context: combined canonical r=0.495±0.052(CI); combined canonical < nature $t(9)=20.9, p<0.001$; combined canonical > transportation $t(9)=60.7, p<0.001$; combined full r=0.545±0.052; combined full < nature $t(9)=16.1, p<0.001$; combined full > transportation $t(9)=67.0, p<0.001$; transportation context: combined canonical r=0.631±0.042(CI); combined canonical > nature $t(9)=7.91, p<0.001$; combined canonical < transportation $t(9)=9.38, p<0.001$; combined full r=0.665±0.032(CI); combined full > nature $t(9)=15.7, p<0.001$; combined full < transportation $t(9)=6.34, p<0.001$).

Crucially, we observed that when using all training examples from one semantic context (e.g., nature, 70M words) and adding new examples from a different context (e.g., transportation, 50M additional words), the resulting embedding space performed worse at predicting human similarity judgments than the contextually-constrained embedding space that used only half of the training data. This result strongly suggests that the contextual relevance of the training data used to generate embedding spaces can be more important than the amount of data itself. Contrary to common practice, adding more training examples may, in fact, degrade performance if the extra training data are not contextually relevant to the relationships of interest (in this case, similarity judgments among items).

Together, these results suggest that combining training data from multiple semantic contexts when generating embedding spaces may be responsible in part for the misalignment between human semantic judgments and the relationships recovered by context-free embedding models (which are usually trained using data from many semantic contexts).

Experiment 2: Contextually-Constrained Embeddings Capture More Information about Interpretable Object Features Compared to Context-Free Embeddings

The findings above are consistent with the hypothesis that semantic judgments are subject to contextual (attentional) influences that constrain the scope of the knowledge representation (e.g., item features) used to make those judgments. However, the analyses presented so far provide little insight into the feature dimensions that define the underlying structure of semantic representations and on which similarity judgments are based. It is largely assumed that, for humans, feature dimensions are recognizable, simple ones, such as size, shape, location, function, etc.[16,23] However, it remains possible that the true underlying representation may instead be comprised of more abstract combinations of such simple features, and/or other perhaps uninterpretable features. From a machine learning perspective, embeddings based on large-scale, unconstrained corpora generally do not yield interpretable features, even when they generate results that capture some aspects of human performance[26,32]. One potential explanation for this may be that contextual constraints play an important role in directing attention to particular features when these are being rated by humans, whereas this contextual influence is weakened when generating context-free embedding spaces (cf. Experiment 1b) and thus information along interpretable features may be less emphasized in such spaces, as well. To evaluate this possibility, we tested whether contextually-constrained embedding spaces would yield feature ratings for individual objects that are more closely aligned to humans on intuitively recognizable dimensions (e.g., size), as well as more relevant to predicting empirical similarity judgments.

To generate feature ratings from embedding spaces, we used an approach based on a method recently proposed by Grand et al.[31]. In this method, for a given feature (e.g., size), a set of three adjectives is chosen that corresponds to the low end of the feature range (e.g., 'small', 'tiny', 'minuscule') and a second set of three adjectives is chosen that corresponds to the high end of the feature range (e.g., 'large', 'huge', 'giant'). The embedding vector for each adjective at one end is then subtracted from the embedding vector of each adjective at the other end (9 differences total). These vector differences are then averaged to create a one-dimensional subspace within the original word embedding space. Other words/concepts are then projected onto that line and the relative distance between each word and the low-/high-end adjective represents a feature rating prediction for that word.

Grand et al.[31] reported a moderate degree of success using this method to predict feature ratings along multiple dimensions in multiple domains. As with the use of word embeddings to predict similarity judgments, the method has been applied only to unconstrained corpora, and thus has ignored the potential influence of

semantic context. However, feature ratings may be impacted by context much as – and perhaps for the same reasons as – similarity judgments (e.g., the size of an ant may be judged to be small among animals, but to be enormous compared to an atom). To test for this, we implemented an extension of the method, in which the endpoints used to identify a dimension are drawn from a particular semantic context rather than the corpus at large. Although Grand et al.[31] referred to their method as "semantic projection", in order to distinguish it from the variant described below, we refer to it as "adjective projection" and to the variant we implemented as "contextual projection".

Contextual projection is similar to adjective projection in that one-dimensional subspaces are used to represent each feature. However, the endpoints of each feature are defined using out-of-sample objects from a given semantic context. For example, rather than evaluating 'size' with respect to a vector from 'small' to 'large' constructed relative to *all* items represented in the embedding space, contextual projection evaluates 'size' in a manner specific to a particular semantic context; that is, using items that are relevant to the context, but not in the set used to test other items (to ensure generality and avoid overfitting). For example, in the nature context, 'size' is measured as the vector from 'mouse' to 'elephant' (*animals* in the training, but not in the testing set) and in the transportation context as the vector from 'skateboard' to 'spaceship' (*vehicles* not in the testing set).

To test how well projection methods predict human feature ratings and similarity judgments, we identified 12 context-relevant features for each of the two semantic contexts used in Experiment 1 (e.g., 'ferocity' for the nature context and 'speed' for the transportation context) and used the Amazon Mechanical Turk platform to collect ratings of those features for the 10 test objects in each associated context; that is, the 10 animals were rated on the 12 nature features and the 10 vehicles were rated on the 12 transportation features (Likert scales 1—5 were used for all features and objects). A full list of features for each semantic context is given in Supplementary Tables 1 & 2. Given the high inter-rater reliability (r=0.68–0.92) observed for test object feature ratings, the contextual projection endpoints were chosen by the experimenters as reasonable examples of additional objects representative of the low/high value on their corresponding feature, but distinct from the 10 test objects used for each semantic context. A complete list of the adjective and contextual projection endpoints used for each semantic context and each feature are listed in Supplementary Tables 3 & 4.

We found that using either adjective or contextual projection we were able to predict human feature ratings reasonably well, suggesting that feature information can be recovered from embedding spaces via projection (Fig. 3). However, we found that contextual projection predicted human feature ratings better than adjective projection. Moreover, for both semantic contexts, using contextually-constrained embedding

spaces together with contextual projection, we were able to predict human feature ratings better than using adjective projection and/or using other embedding spaces for 10 out of 12 context-relevant features (Fig. 3). These results were not sensitive to the initialization conditions of the embedding models used for predicting feature ratings (Fig. 3 includes 95% confidence intervals for 10 independent initializations of each model).

Nature Context

Dimension / Model	Aquaticness	Cuteness	Dangerousness	Domesticity	Edibility	Furriness	Humanness	Intelligence	Interest	Predacity	Size	Speed
Contextual Projection												
Nature	**.63 ± .01**	.16 ± .07	.42 ± .07	**.35 ± .09**	**.59 ± .06**	**.48 ± .01**	**.50 ± .05**	**.41 ± .12**	**.35 ± .05**	**.43 ± .07**	**.86 ± .06**	**.43 ± .14**
Transportation	.21 ± .19	.04 ± .07	.10 ± .12	.11 ± .11	.02 ± .07	.02 ± .04	.02 ± .04	.08 ± .21	**.46 ± .32**	.23 ± .15	.04 ± .07	.27 ± .11
Wikipedia (full)	.37 ± .03	**.42 ± .06**	.25 ± .03	.11 ± .07	**.46 ± .08**	**.41 ± .03**	.20 ± .04	.01 ± .02	.03 ± .02	**.37 ± .03**	.28 ± .11	**.46 ± .06**
Adjective Projection												
Nature	**.71 ± .15**	.23 ± .12	.29 ± .19	**.48 ± .30**	.33 ± .11	**.36 ± .24**	.23 ± .14	.08 ± .09	**.15 ± .29**	**.34 ± .27**	.31 ± .27	.27 ± .21
Transportation	.23 ± .05	.24 ± .09	**.75 ± .09**	.06 ± .03	.04 ± .04	.11 ± .01	.00 ± .01	.08 ± .04	**.19 ± .05**	.23 ± .12	.08 ± .05	.01 ± .03
Wikipedia (full)	.45 ± .13	.02 ± .07	.19 ± .26	.22 ± .22	.18 ± .11	.05 ± .06	.05 ± .06	**.37 ± .10**	**.19 ± .10**	.14 ± .09	.26 ± .24	.03 ± .09

Transportation Context

Dimension / Model	Comfort	Cost	Dangerousness	Elevation	Interest	Openness	Personalness	Size	Skill	Speed	Usefulness	Wheeledness
Contextual Projection												
Nature	**.08 ± .06**	.77 ± .07	**.67 ± .04**	.36 ± .06	**.68 ± .06**	**.65 ± .04**	.72 ± .05	**.65 ± .06**	.70 ± .05	**.36 ± .17**	**.12 ± .16**	**.90 ± .03**
Transportation	**.16 ± .11**	**.89 ± .06**	**.73 ± .15**	**.54 ± .08**	.34 ± .06	**.70 ± .08**	.67 ± .13	**.75 ± .04**	**.82 ± .03**	**.48 ± .09**	**.17 ± .01**	**.87 ± .03**
Wikipedia (full)	**.12 ± .05**	**.85 ± .02**	.55 ± .02	.36 ± .02	**.64 ± .03**	.57 ± .02	**.84 ± .02**	**.71 ± .03**	.74 ± .03	.26 ± .09	.04 ± .03	**.88 ± .02**
Adjective Projection												
Nature	**.34 ± .24**	.21 ± .24	.04 ± .11	.09 ± .04	.25 ± .17	.01 ± .04	.27 ± .35	.25 ± .30	.42 ± .13	.06 ± .02	.01 ± .03	.78 ± .09
Transportation	**.06 ± .09**	.24 ± .13	.39 ± .07	.03 ± .03	.11 ± .12	.03 ± .05	.34 ± .21	.06 ± .08	.60 ± .09	**.35 ± .23**	**.25 ± .07**	.84 ± .02
Wikipedia (full)	**.19 ± .22**	.18 ± .21	.02 ± .05	.05 ± .08	.02 ± .04	.04 ± .09	.18 ± .21	.10 ± .17	.19 ± .07	**.47 ± .24**	.10 ± .03	**.84 ± .09**

Fig. 3. Contextual and Adjective Projection Recover Human Feature Ratings. *Pearson correlations between predicted feature ratings using the contextual and adjective projection methods for items in the nature context (animals) and items in the transportation context (vehicles) with empirically obtained human feature ratings for corresponding semantic contexts. For both nature and transportation semantic contexts, using contextual projection in the contextually-constrained embedding spaces generated ratings that were better aligned with human judgments compared to other models and projection methods for 10 out of the 12 features considered. Errors signify 95% confidence intervals for 10 independent initializations of the model training procedure. Bolding and highlights indicate best (or tied for best) performing model in each column (red – contextually-constrained nature; blue – contextually-constrained transportation; green – context-free).*

Similar projection procedures applied to the canonical combined-context embedding space generated in Experiment 1b (50% nature–50% transportation, 60M words) yielded feature information that was less well aligned with human feature ratings, compared to the pure contextually-constrained models, but better aligned than the contextually-constrained models for the other context, or the context-free models (Fig. 4).

Nature Context

Dimension / Model	Aquaticness	Cuteness	Dangerousness	Domesticity	Edibility	Furriness	Humanness	Intelligence	Interest	Predacity	Size	Speed
Contextual Projection												
Nature	**.63 ± .01**	.16 ± .07	**.42 ± .07**	**.35 ± .09**	**.59 ± .06**	**.48 ± .01**	**.50 ± .05**	**.41 ± .12**	**.35 ± .05**	**.43 ± .07**	**.86 ± .06**	**.43 ± .14**
Combined	.42 ± .25	**.57 ± .11**	**.35 ± .17**	**.50 ± .10**	.02 ± .05	.32 ± .12	**.51 ± .12**	.22 ± .14	.13 ± .06	.26 ± .11	.60 ± .13	**.37 ± .08**
Transportation	.21 ± .19	.04 ± .07	.10 ± .12	.11 ± .11	.02 ± .07	.02 ± .04	.02 ± .04	.08 ± .21	**.46 ± .32**	.23 ± .15	.04 ± .07	.27 ± .11
Wikipedia (full)	.37 ± .03	**.42 ± .06**	.25 ± .03	.11 ± .07	**.46 ± .08**	**.41 ± .03**	.20 ± .04	.01 ± .02	.03 ± .02	**.37 ± .03**	.28 ± .11	**.46 ± .06**

Transportation Context

Dimension / Model	Comfort	Cost	Dangerousness	Elevation	Interest	Openness	Personalness	Size	Skill	Speed	Usefulness	Wheeledness
Contextual Projection												
Nature	.08 ± .06	.77 ± .07	**.67 ± .04**	.36 ± .06	**.68 ± .06**	.65 ± .04	.72 ± .05	**.65 ± .06**	.70 ± .05	**.36 ± .17**	**.12 ± .16**	**.90 ± .03**
Combined	**.83 ± .04**	.07 ± .07	**.76 ± .06**	**.53 ± .08**	.34 ± .10	**.79 ± .03**	.46 ± .12	.36 ± .09	.66 ± .08	.20 ± .15	.02 ± .04	.54 ± .07
Transportation	.16 ± .11	**.89 ± .06**	**.73 ± .15**	**.54 ± .08**	.34 ± .06	**.70 ± .08**	.67 ± .13	**.75 ± .04**	**.82 ± .03**	**.48 ± .09**	**.17 ± .01**	**.87 ± .03**
Wikipedia (full)	.12 ± .05	**.85 ± .02**	.55 ± .02	.36 ± .02	**.64 ± .03**	.57 ± .02	**.84 ± .02**	**.71 ± .03**	.74 ± .03	.26 ± .09	.04 ± .03	**.88 ± .02**

Fig. 4. Combined-Context Embedding Spaces Recover Feature Ratings Less Well Than Contextually-Constrained Embedding Spaces. *Pearson correlation between predicted feature ratings using contextual projection applied to the canonical combined-context embedding space (50% nature – 50% transportation, 60M words) and empirical human feature ratings. In the nature context, the contextually-constrained embeddings were best aligned with human judgments on 11 out of the 12 features considered and the combined-context embeddings were tied for best for 4 out of 12 features and performed best for only one feature (cuteness). In the transportation context, the contextually-constrained embeddings were best aligned with human judgments for 9 out of the 12 features considered and the combined-context embeddings were tied for best for 3 out of 12 features and performed best for only one feature (comfort). Errors signify 95% confidence intervals estimated from 10 independent initializations of the model training procedure. Bolding and highlights indicate best (or tied for best) performing model in each column (red – contextually-constrained nature; purple – context-combined; blue – contextually-constrained transportation; green – context-free).*

Together, the findings of Experiments 1 and 2 support the hypothesis that, using projection methods, contextually-constrained embedding spaces can predict human features ratings better than context-free embedding spaces in their respective contexts. We also show that training embedding spaces on corpora that include multiple semantic contexts substantially degrades both the ability to predict object-level similarity judgments (Experiment 1), as well as object feature values (Experiment 2), even though these types of judgments are easy for humans to make and reliable across individuals.

Experiment 3: Using Contextually-Relevant Features to Improve Prediction of Human Similarity Judgments from Context-Free Embeddings

Context-free embeddings are built from large-scale corpora comprising billions of words that likely span hundreds of semantic contexts. Currently, such embedding spaces are a key component of many application domains, ranging from neuroscience[8,9] to computer science[10,11,12,13] and beyond[14]. Our work suggests that if the goal of these applications is to solve human-relevant problems, then at least some of these domains may benefit from employing contextually-constrained embedding spaces instead, which would better predict human semantic structure. However, re-training embedding models using different text corpora and/or collecting such semantically-relevant corpora on a case-by-case basis may be expensive or difficult in practice. To help alleviate this problem, we propose an alternative approach that uses contextually-relevant features to predict human semantic information (e.g., similarity judgments) from *context-free* embedding spaces.

Previous work in cognitive science has attempted to predict similarity judgments from object feature values by collecting empirical ratings for objects along different features and computing the distance (using various metrics) between those feature vectors for pairs of objects. Such methods consistently explain about a third of the variance observed in human similarity judgments[3,4,16,21,41]. They can be further improved using regression methods to differentially weight the dimensions, but at best this can only explain about half the variance in human similarity judgments (e.g., $r=0.65$[23]).

Here, we test the hypothesis that human similarity judgments can be better predicted from context-free embedding spaces by using contextually-relevant features (cf. Experiment 2) together with the regression methods employed in cognitive psychology experiments that attempt to predict similarity between objects based on such features. First, we used the contextual projection method described in Experiment 2 to construct 12-dimensional subspaces for each embedding space

corresponding to the 12 features in each semantic context (see Supplementary Tables 1 & 2 for details on features and endpoints). Second, we used linear regression to learn optimal weights for each feature in the 12-dimensional subspace of each embedding space that together best predicted human similarity judgments. We then performed out of sample prediction by using cross-validation to repeatedly select 80% of the empirical data for learning the weights and predict similarity using the learned weights on the left-out 20% of the judgments (Fig. 5).

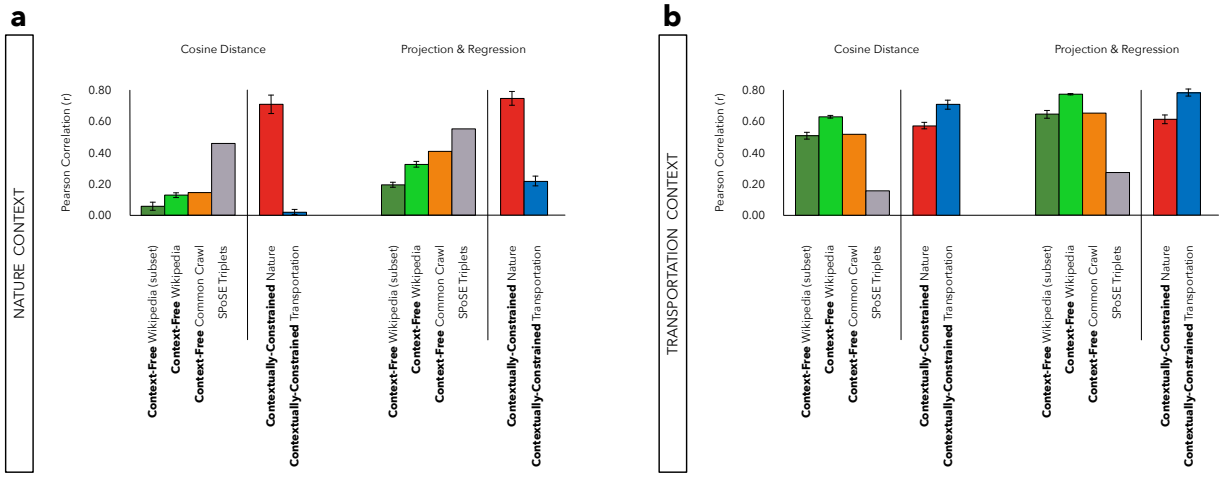

Figure 5. Contextual Projection and Linear Regression Significantly Improve Recovery of Human Similarity Judgments from Embedding Spaces. *Contextual projection was used to generate 12-dimensional subspaces for each embedding space corresponding to the 12 features for each semantic context. Linear regression was then used to learn optimal weights for each feature in each subspace that together best predicted human similarity judgments. Graphs show Pearson correlation between out-of-sample cross-validated predicted similarity values using the projection and regression procedure and human similarity judgments. (a) Nature semantic context. (b) Transportation semantic context. Errors signify 95% confidence intervals for 10 independent initializations of the word embedding models' training procedure. All differences between identically colored bars are statistically significant, p<0.001.*

The contextual projection and regression procedure significantly improved predictions of human similarity judgments for all context-free embedding spaces (Fig. 5; nature context, projection & regression > cosine: Wikipedia subset $t(9)=24.1, p<0.001$; Wikipedia $t(9)=105, p<0.001$; Common Crawl $r=0.409>r=0.145$; SPoSE $r=0.554>r=0.460$; transportation context, projection & regression > cosine: Wikipedia subset $t(9)=22.6, p<0.001$; Wikipedia $t(9)=132, p<0.001$; Common Crawl $r=0.654 > r=0.519$; SPoSE $r=0.273 > r=0.156$). By comparison, neither learning weights

on the original set of 100 dimensions in each embedding space via regression (Supplementary Fig. 5), nor using cosine distance in the 12-dimensional contextual projection space, which is equivalent to assigning the same weight to each feature (Supplementary Fig. 6), could predict human similarity judgments as well as using both contextual projection and regression together. These results suggest that the improved accuracy of combined contextual projection and regression provide a novel and more accurate approach for recovering human-aligned semantic relationships that appear to be present, but previously inaccessible, within context-free embedding spaces.

Finally, if people differentially weight different dimensions when making similarity judgments, then the contextual projection and regression procedure should also improve predictions of human similarity judgments from contextually constrained embeddings. Our findings not only confirm this prediction (Fig. 5; nature context, projection & regression > cosine: nature $t(9)=8.79, p<0.001$, transportation $t(9)=22.8, p<0.001$, Wikipedia $t(9)=105, p<0.001$; transportation context, projection & regression > cosine: nature $t(9)=15.0, p<0.001$, transportation $t(9)=42.5, p<0.001$, Wikipedia $t(9)=132, p<0.001$), but also provide the best prediction of human similarity judgments to date using either human feature ratings or text-based embedding spaces, with correlations of up to $r=0.75$ in the nature semantic context and up to $r=0.78$ in the transportation semantic context. This accounts for 92% and 90% of human inter-rater variability in human similarity judgments for these two contexts. Contextual projection with learned weights shows substantial improvement upon the best previous prediction of human similarity judgments using empirical human feature ratings ($r=0.65$[23]). Remarkably, the predictions are from features extracted from artificially-built word embedding spaces, not empirical human feature ratings, and are evaluated using an out-of-sample prediction procedure. The ability to reach or exceed 90% of human inter-rater reliability in these specific semantic contexts suggests that this computational approach provides an accurate and robust representation of the structure of human semantic knowledge.

Discussion

The results of the three experiments we reported support the hypothesis that context plays a key role in human semantic processing, and that efforts to understand how semantic knowledge is organized can benefit by taking context into account, whether in empirical studies or when using machine learning methods applied to large-scale text corpora. Moreover, we showed that such machine learning approaches can be used to reliably estimate human semantic similarity judgments and object

feature ratings, provided that information relevant to the contextual effects of human attention is used, either in the construction of the training corpora (Experiments 1 and 2) or in the method used to make predictions about human judgments from embedding spaces (Experiment 3).

From a psychological and cognitive science perspective, discovering reliable mappings between data-driven approaches and human judgments may help improve long-standing models of human behavior for tasks such as categorization, learning, and prediction. Understanding how people carry out such tasks requires the ability to reliably estimate similarity between concepts, identify features that describe them, and characterize how attention impacts these measurements, efforts that, for practical reasons, have so far focused on either artificially-built examples or small-scale subsets of cognitive space[2,3,21,23,35,42]. To the extent that the improvement in alignment evidenced by our results between artificial semantic spaces (embeddings) and human semantic structure can be further validated, it will allow us to study these tasks and the influence of attention on them much more efficiently and at a much broader scale than is currently possible in laboratory studies alone.

From a human neuroscience perspective, it is unlikely that humans retrain their long-standing semantic representations every time a new task demands it; instead attention is thought to alter the context in which learned semantic structure is processed for the purpose at hand, both in behavior and in the brain[43,44,45,46]. Recent advances in neuroimaging have allowed embedding-based neural models of semantics to probe how concepts are processed across the human brain[8,47] and to generate decoders of mental representations that can predict human behavior from neural responses[9]. Increasing alignment between embedding spaces and human semantic structure helps further our understanding of the structural underpinnings of semantic knowledge. Together with recent neuroimaging work that suggests an interaction between cognitive control (anterior cingulate cortex, ACC) and infero-temporal cortex while humans perform similarity judgments[18,48], our results also suggest a novel avenue for investigating the mechanisms of how context dynamically shifts human behavior and neural responses across large-scale semantic structure.

From a machine learning perspective, embedding spaces have been used extensively in natural language processing (NLP) tasks as a primary building block, under the assumption that these spaces represent adequate proxies for human syntactic and semantic structure. By substantially improving alignment of these embeddings with empirical object feature ratings and similarity judgments, we introduce new tools for exploration in NLP. Both human-aligned embedding spaces resulting from contextually-constrained training, and projections that are motivated and validated on a small amount of empirical data, may lead to improvements in the performance of NLP models that rely on embedding spaces to make inferences about

human decision making and task performance. Example applications include machine translation[10], automatic extension of knowledge bases[11], text summarization[12], image and video captioning[49,50,51,52].

In NLP (and more broadly in machine learning), it is generally assumed that increasing the size of the training corpus should increase performance[26,30]. However, our results suggest an important countervailing factor: the extent to which the training corpus reflects the influence of the same attentional factors (context) as the subsequent testing regime. In our experiments, contextually-constrained models trained on corpora comprising 50-70 million words outperformed state-of-the-art context-free models trained on billions of words. Furthermore, building a combined-context model using all 120 million words from both the nature and transportation contexts yielded an embedding space that did not improve on the alignment with human judgments compared to either of the (smaller) contextually-constrained models in their preferred contexts (Figs. 2–4). These findings demonstrate that context is a key factor in building embedding spaces that are aligned with human semantic space and that, more broadly, data quality (as measured by contextual relevance) may be just as important as data quantity in building embedding spaces that would capture relationships salient to the specific task in which such spaces are employed.

The method and observations we report strengthen the existing link between human semantic space (how we organize knowledge and use it to interact with the world) and machine learning methods meant to automate tasks useful and directly relevant to humans (e.g., NLP). The ability to reach or exceed 90% of maximum achievable performance in predicting similarity between concepts in these specific semantic contexts validates a set of computational tools that allow for more accurate and robust representations of human semantic knowledge and are likely to be helpful in the future in understanding the underlying structure of human semantic representations and also in efforts to build artificial systems that can emulate and/or better interact with semantic representations.

Methods

Generating Word Embedding Spaces

We generated all semantic embedding spaces using the continuous skip-gram Word2Vec model with negative sampling as proposed by Mikolov et al.[26,27], henceforth referred to as 'Word2Vec'. Word2Vec has been shown to be superior to, or on par with, other embedding models at matching human similarity judgments[30]. The assumption behind Word2Vec is that words that appear in similar syntactic contexts

(i.e., surrounded by, co-occurring with, or in a window size of a similar set of 8–12 words) tend to have similar meanings. To encode this relationship between words, Word2Vec represents each word as a multidimensional vector. The algorithm preferentially learns word vectors of a specific dimension such that the ability of a vector representation of a word to predict other words within a given syntactic window is maximized (i.e., words from the same syntactic window are placed close to each other in this multidimensional space, as are words whose syntactic windows are similar to one another).

We trained four primary types of embedding spaces: (1) contextually-constrained models ('nature' and 'transportation'); (2) context-combined models; and (3) context-free models. Contextually-constrained models (1) were trained on a subset of English language Wikipedia determined by human-curated category labels (meta-information available directly from Wikipedia) associated with each Wikipedia article. Each category contains multiple articles and multiple sub-categories; the categories of Wikipedia thus form a tree where articles themselves are the leaves. We constructed the 'nature' semantic context training corpus by traversing the tree rooted at the 'animal' category, and we constructed the 'transportation' semantic context training corpus by combining the trees rooted at the 'transport' and 'travel' categories. To avoid topics unrelated to natural semantic contexts, we removed the sub-tree 'humans' from the 'nature' training corpus. Furthermore, to ensure that the animal and vehicle contexts were non-overlapping, we removed training articles that were labeled as belonging to both the 'nature' and 'transportation' training corpora. This yielded final training corpora of approximately 70 million words for the 'nature' semantic context and 50 million words for the 'transportation' semantic context. The combined-context models (2) were trained by combining data from each of the two contextually-constrained training corpora in varying amounts. For the models that matched training corpora size with the contextually-constrained models, we selected proportions of the two corpora that added up to approximately 60 million words (e.g., 10% 'transportation' corpus + 90% 'nature' corpus). The canonical size-matched combined-context model was obtained using a 50%-50% split (i.e., approximately 35 million words from the 'nature' semantic context and 25 million words from the 'transportation' semantic context). We also trained a combined-context model that included all training data used to generate both the 'nature' and the 'transportation' contextually-constrained models (full combined-context model, approximately 120 million words). Finally, the context-free models (3) were trained using English language Wikipedia articles unrestricted to a particular category (or semantic context). The full context-free Wikipedia model was trained using the full corpus of text corresponding to all Wikipedia articles (approximately 2 billion words) and the size-matched context-free model was trained by sampling 60 million words from this full corpus.

The primary factors controlling the Word2Vec model are the size of the syntactic context window ('window size') and the dimensionality of the resulting embedding space. Larger window sized result in embedding spaces that capture relationships between words that are farther apart in a document, and larger dimensionality has the potential to represent more of these relationships between words in a vocabulary. In practice, as window size or vector length increase, larger amounts of training data are required. To build our embedding spaces, we first conducted a grid search of all window sizes in the set (8, 9, 10, 11, 12) and all dimensionalities in the set (100, 150, 200) and selected the combination of parameters which yielded the highest agreement between similarity predicted by the Wikipedia context-free model (4) (2 billion words) and empirically-collected human similarity judgments (see *Human Behavioral Experiments* section). We reasoned that this would provide the most stringent possible benchmark of the context-free embedding spaces for evaluating our contextually-constrained embedding spaces against. Accordingly, all results and figures in the manuscript are obtained using models with a window size of 9 words and a dimensionality of 100 (Supplementary Figs. 1 & 2).

All models were trained using the 'gensim' Python library's implementation of the Word2Vec model[53]. Aside from window size and dimensionality, all other parameters were kept as the default values from the original Word2Vec publications[26,27]: an initial learning rate of 0.025, elimination of words that appear fewer than 5 times in the training corpus, a 0.001 threshold for downsampling frequently occurring words, an exponent of 0.75 for shaping the negative sampling distribution, 5 negative samples per positive sample, and the skip-gram training algorithm. The resulting vocabulary sizes for each embedding space we constructed were: 148K vectors for the contextually-constrained 'nature' model, 110K words for the contextually-constrained 'transportation' model, 204K words for the combined-context models (canonical & full), 342K words for the context-free Wikipedia full model, and 125K words for the context-free Wikipedia subset model.

For each type of model (contextually-constrained, combined-context, context-free) we trained ten separate models with different initializations (but identical parameters) to control for the influence that random initialization of the weights has on model performance. Cosine similarity was used as a distance metric between two learned word vectors. Results from different initializations are included as 95% confidence intervals in main text and supplementary figures (Figs. 1–5, Supplementary Figs. 1–2, 5–6).

We also compared against the pre-trained GloVe embedding space[28] generated using a corpus of 42 billion words (freely available online: https://nlp.stanford.edu/projects/glove/). The pre-trained GloVe model had a dimensionality of 300 and a vocabulary size of 400K words.

Finally, we compared the performance of our context-specific embedding spaces against a recent large-scale concept similarity model based on estimating ordinal similarity between triplets of objects (SPoSE[39]). We compared against this dataset as it represents the largest scale attempt to date to predict human similarity judgments in any form and because it generates similarity predictions for all the test objects we selected in our study (all pairwise comparisons between our test stimuli shown below are included in the output of the SPoSE model).

Object and Feature Testing Sets

To test how well the trained embedding spaces aligned with human cognitive judgments, we constructed a stimulus set comprising ten representative basic-level animals (bear, cat, deer, duck, parrot, seal, snake, tiger, turtle, and whale) for the naturen semantic context and ten representative basic-level vehicles (airplane, bicycle, boat, car, helicopter, motorcycle, rocket, shuttle, submarine, truck) for the transportation semantic context (Fig. 1b). We also selected twelve human-relevant features independently for each semantic context which have been shown to explain object-level similarity judgments in empirical settings[3,23,54]. These features are aquaticness, cuteness, dangerousness, domesticity, edibility, furriness, humanness, intelligence, interesting-ness, predacity, size, speed for the nature semantic context and comfort, cost, dangerousness, elevation, interesting-ness, open-ness, personal-ness, size, skill, speed, usefulness, wheeled-ness for the transportation semantic context (Table S1).

For each of the twenty total object categories (e.g., bear (animal), airplane (vehicle)), we collected nine images showcasing the animal in its natural habitat or the vehicle in its normal domain of operation. All images were in color, featured the target object as the largest and most prominent object on the screen, and were cropped to a size of 500x500 pixels each (one representative image from each category shown in Fig. 1b).

Human Behavioral Experiments

To collect empirical similarity judgments, we recruited 139 participants (45 female, 108 right-handed, mean age 31.5 years) through the Amazon Mechanical Turk online platform in exchange for $1.50 payment (expected rate $7.50/hour). Participants were asked to report the similarity between every pair of objects from a single semantic context (e.g., all pairwise combinations of ten vehicles or all pairwise combinations of ten animals) on a discrete scale of 1 to 5 (1 = not similar; 5 = very similar). In each trial, the participant was shown two randomly selected images from

each category side-by-side and was given unlimited time to report a similarity judgment. Each participant made 45 comparisons (all pairwise combinations of 10 categories from a single randomly chosen semantic context) presented in a random order.

To ensure high-quality judgments, we limited participation only to Mechanical Turk workers who had previously completed at least 1000 HITs with an acceptance rate of 95% or above. We excluded 34 participants who had no variance across answers (e.g. choosing a similarity value of 1 for every object pair). Prior work has shown that for this type of task inter-participant reliability should be high[23], therefore to exclude participants whose response may have been random, we correlated the responses of each participant with the average of the responses for every other participant and calculated the Pearson correlation coefficient. We then iteratively removed the participant with the lowest Pearson coefficient, stopping this procedure when all remaining participants had a Pearson coefficient greater than or equal to 0.5 to the rest of the group. This excluded an additional 12 participants, leading to a final tally of n=44 participants for the nature semantic context and n=49 participants for the transportation semantic context.

To collect empirical feature ratings, we recruited 915 participants (392 female, 549 right-handed, mean age 33.4 years) through the Amazon Mechanical Turk online platform in exchange for $0.50 payment (expected rate $7.50/hour). Participants were asked to rank every object from a single semantic context (e.g., all ten vehicles or all ten animals) along a randomly chosen, context-specific dimension (e.g., "How fast/slow is this vehicle?") on a discrete scale of 1 to 5 (1 = low feature value, e.g. 'slow'; 5 = high feature value, e.g. 'fast'). In each trial, the participant was shown three randomly selected images from a total of nine possible images representing the object, as well as the name of the object (e.g., 'bear') and given unlimited time to report a feature rating. Each participant ranked all ten objects, presented in a random order, from a single randomly chosen context along a single randomly chosen dimension.

We used an analogous procedure as in collecting empirical similarity judgments to select high-quality responses (e.g., restricting the experiment to high performing workers and excluding 210 participants with low variance responses and 124 participants with answers that correlated poorly with the average response). This resulted in 18–33 total participants per feature (see Supplementary Tables 1 & 2 for details).

All participants had normal or corrected-to-normal visual acuity and provided informed consent to a protocol approved by the Princeton University Institutional Review Board.

Predicting Similarity Judgments from Embedding Spaces

To predict similarity between two objects in an embedding space, we computed the cosine distance between the vector representations of each object. We used cosine distance as a metric for two main reasons. First, cosine distance is a commonly reported metric used in the literature that allows for direct comparison to previous work[26,27,28,30]. Second, cosine distance disregards the length or magnitude of the two vectors being compared, taking into account only the angle between the vectors. Some studies[55] have demonstrated a relationship between the frequency with which a word appears in the training corpus and the length of the word vector. Because this frequency relationship should not have any bearing on the semantic similarity of the two words, using a distance metric such as cosine distance that ignores magnitude/length information is prudent.

Additionally, we developed a novel method for predicting semantic similarity judgments from word embedding spaces by first projecting embeddings onto human-relevant features (see section on *Feature Projections in Embedding Spaces* below) and then using these feature ratings to determine semantic similarity. To do so, we used either adjective or contextual projection to generate ratings for objects along the twelve human-relevant features we selected for each semantic context and subsequently used linear regression to estimate an optimal weighting for the features in each embedding space that would yield the best prediction of human similarity judgments. We used a leave-one-object-out cross-validation procedure where all pairwise comparisons between 9 of the 10 objects in each semantic context were used to estimate the feature weights and then the weights were used to predict similarity between the left-out object and the other 9 objects used for training.

Feature Projections in Embedding Spaces

To generate putative ratings for objects along particular features using embedding spaces, we adapted the projection method developed by Grand et al.[28]. The authors dubbed their technique 'semantic projection', but to avoid confusion with our own work, we henceforth refer to it as 'adjective projection'.

Adjective projection starts with manually defining three adjectives each that represent the extreme ends of a particular feature (e.g., for the "size" feature, adjectives representing the low end are 'small', 'tiny', and 'little', and adjectives representing the high end are 'large', 'huge', and 'giant'). Subsequently, for each feature, nine new vectors are constructed in the embedding space as the vector differences between all possible pairs of an adjectives representing the low extreme of a feature and an adjective representing the high extreme of a feature (e.g., the vector difference

between line 'small' and line 'large', line 'small' and line 'huge', etc.). The average of these nine vectors represents a one-dimensional subspace of the original embedding space and is used as approximation of its corresponding feature (e.g., the 'size' feature vector). All adjectives used across each feature and semantic context are shown in Supplementary Table 1 (nature context) and Supplementary Table 2 (transportation context).

Once a feature subspace was defined, the rating of an object with respect to that feature was determined by projecting the vector representing the object in the original embedding space onto the one-dimensional feature subspace, which resulted in a scalar value (overall range across all models, features, and contexts: [-0.6, 0.4]):

$$rating_{object} = \frac{feature^T object}{||feature||}$$

To illustrate the relationship with cosine distance in the original embedding space, we note that the difference between the feature ratings of two words is then equivalent to the normalized cosine distance between the vector difference of those two words in the original embedding space and the corresponding feature vector:

$$dist(object_1, object_2) = \frac{feature^T(object_1 - object_2)}{\left|\left|feature\right|\right|} =$$

$$= cosineDist(object_1 - object_2, feature) \cdot ||object_1 - object_2||$$

To take advantage of our insight that semantic context information is key to recovering human-relevant information from embedding spaces, we extended the work of Grand et al.[28] by defining a new contextual projection method. Contextual projection generates feature ratings using an analogous procedure to adjective projection, except for the fact that the low and high endpoints of each feature are defined by objects relevant to a particular semantic context, rather than adjectives common to all semantic contexts. For example, for the 'size' feature in the nature semantic context, we selected 'bird', 'rabbit', 'rat' as low endpoints and 'lion', 'giraffe', 'elephant' as high endpoints, instead of 'small', 'tiny', 'little', and 'large', 'huge', 'giant', respectively. To maximize the amount of human data available for testing the performance of our embedding spaces (and given the high inter-rater reliability, r=0.68–0.92, observed for test object feature ratings), we chose as endpoints for the contextual projection examples of objects that did not overlap with the testing sets in each semantic context. A complete list of the adjective and contextual projection

endpoints used for each semantic context and each feature are listed in Supplementary Table 3 (nature context) and Table Supplementary 4 (transportation context).

Statistics

All error bars reported are 95% confidence intervals with n=10 different learned embedding representations (e.g., 10 different initial conditions), degrees of freedom = 9. All correlation values reported are Pearson r correlation coefficients. All t-tests performed were paired and two-tailed, n=10, degrees of freedom = 9.

References

1. Tversky, A. Features of similarity. Psych. Rev. **84,** 327-352 (1977).
2. Nosofsky, R. M. Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. Perception & Psychophys. **38,** 415-432 (1985).
3. Osherson, D. N., Stern, J., Wilkie, O., Stob, M. & Smith, E. E. Default Probability. Cogn. Sci. **15,** 251–269 (1991).
4. Rogers, T. T. & McClelland, J. L. *Semantic cognition: a parallel distributed processing approach*. (MIT Press, 2004).
5. Nosofsky, R. M. Attention, similarity, and the identification-categorization relationship. J. Exp. Psychol. Gen. **115,** 39–57 (1986).
6. Smith, E. E. & Medin, D. L. *Categories and concepts*. (Harvard University Press, 1981).
7. Murphy, G. L. *The big book of concepts* (MIT Press, 2002).
8. Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature **532,** 453–458 (2016).
9. Pereira, F. et al. Toward a universal decoder of linguistic meaning from brain activation. Nat. Commun. **9,** (2018).
10. Mikolov, T., Yih, S. W. & Zweig, G. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 746-751 (2013).
11. Toutanova, K. et al. Representing text for joint embedding of text and knowledge bases. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing 1499–1509 (2015).
12. Rossiello, G., Basile, P., & Semeraro, G. Centroid-based text summarization through compositionality of word embeddings. In Proceedings of the MultiLing 2017

Workshop on Summarization and Summary Evaluation Across Source Types and Genres, 12-21 (2017).
13. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information. Trans. Association Comput. Linguist. **5,** 135-146 (2017).
14. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. Science **80,** 183–186 (2017).
15. Ashby, F. G. & Lee, W. W. Predicting similarity and categorization from identification. J. Exp. Psychol. Gen. **120,** 150–72 (1991).
16. Nosofsky, R. M. Tests of an exemplar model for relating perceptual classification and recognition memory. J. Exp. Psychol. Hum. Percept. Perform. **17,** 3–27 (1991).
17. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. Trends Cogn. Sci. **11,** 333–341 (2007).
18. Lambon Ralph, M. A., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. Nat. Rev. Neurosci. **18,** 42–55 (2017).
19. Collins, A. M. & Loftus, E. F. A spreading-activation theory of semantic processing. Psychol. Rev. **82,** 407–428 (1975).
20. Rips, L. J. Similarity, typicality, and categorization. in Similarity and Analogical Reasoning (eds. Vosniadou, S. & Ortony, A.) 21–59 (Cambridge University Press, 1989).
21. Maddox, W. T. & Ashby, F. G. Comparing decision bound and exemplar models of categorization. Percept. Psychophys. **53,** 49–70 (1993).
22. Gentner, D. & Markman, A. B. Structural alignment in comparison: No difference without similarity. Psychol. Sci. **5,** 152–158 (1994).
23. Iordan, M. C., Ellis, C. T., Lesnick, M., Osherson, D. N. & Cohen, J. D. Feature ratings and empirical dimension-specific similarity explain distinct aspects of semantic similarity judgments. In Proceedings of the 40th Annual Conference of the Cognitive Science Society, 530-535 (2018).
24. Gentner, D. & Markman, A. B. Structural alignment in comparison: No difference without similarity. Psychol. Sci. **5,** 152–158 (1994).
25. Biederman, I. Recognition-by-components: A theory of human image understanding. Psychol. Rev. **94,** 115–147 (1987).
26. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Advances in Neural Information Processing Systems (NIPS), 3111-3119 (2013).
27. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient Estimation of Word Representations in Vector Space. Preprint at arxiv.org/pdf/1301.3781.pdf (2013).

28. Pennington, J., Socher, R. & Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543 (2014).
29. Baroni, M., Dinu, G. & Kruszewski, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics **1,** 238–247 (2014).
30. Pereira, F., Gershman, S., Ritter, S. & Botvinick, M. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. Cogn. Neuropsychol. **33,** 175–190 (2016).
31. Grand, G., Blank, I. A., Pereira, F. & Fedorenko, E. Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings. Preprint at *arxiv.org/pdf/1802.01241.pdf* (2018).
32. Richie, R., Zou, W. & Bhatia, S. Semantic representations extracted from large language corpora predict high-level human judgement in seven diverse behavioral domains. Preprint at osf.io/vpucz (2019).
33. Gentner, D. Why nouns are learned before verbs: linguistic relativity versus natural partitioning. Lang. Dev. **2,** 301–334 (1982).
34. Dillard, J. P., Palmer, M. T. & Kinney, T. A. Relational judgements in an influence context. Hum. Commun. Res. **21,** 331–353 (1995).
35. Goldstone, R. L., Medin, D. L. & Halberstadt, J. Similarity in context. Mem. Cognit. **25,** 237–255 (1997).
36. Medin, D. L. & Schaffer, M. M. Context theory of classification learning. Psychol. Rev. **85,** 207–238 (1978).
37. Miller, G. A. & Charles, W. G. Contextual correlates of semantic similarity. Lang. Cogn. Process. **6,** 1–28 (1991).
38. Nosofsky, R. M. Choice, similarity, and the context theory of classification. J. Exp. Psychol. Learn. Mem. Cogn. **10,** 104–114 (1984).
39. Zheng, C. Y., Pereira, F., Baker, C. I. & Hebart, M. N. Revealing interpretable object representations from human behavior. Preprint at arxiv.org/pdf/1901.02915.pdf (2019).
40. Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. & Boyes-Braem, P. Basic objects in natural categories. Cogn. Psychol. **8,** 382-439 (1976).
41. Tversky, B. & Hemenway, K. Objects, parts, and categories. J. Exp. Psychol. Gen. **113,** 169–193 (1984).
42. Nosofsky, R. M., Sanders, C. A., & McDaniel M. A. Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. J. Exp Psychol.: Gen. **147,** 328-353 (2018).
43. Rosch, E. & Lloyd, B. L. *Cognition and categorization*. (Lawrence Erlbaum, 1978).

44. Miller, E. K., & Cohen, J. D. An integrative theory or prefrontal cortex function. Annu. Rev. Neurosci. **24,** 167-202 (2001).
45. Bar, M. Visual objects in context. Nat. Rev. Neurosci. **5,** 617–629 (2004).
46. Çukur, T., Nishimoto, S., Huth, A. G. & Gallant, J. L. Attention during natural vision warps semantic representation across the human brain. Nat. Neurosci. **16,** 763–770 (2013).
47. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron **76,** 1210–24 (2012).
48. Keung, W., Osherson, D. N., & Cohen, J. D. Influence of cognitive control on semantic representation. Preprint at bioRXiv.org/content/bioRXiv/early/2016/08/22/067553/full.pdf (2016).
49. Kiros, R., Salakhutdinov, R. & Zemel, R. Multimodal neural language models. in Proceedings of the 31st International Conference on Machine Learning, 595–603 (2014).
50. Hendricks, L.A., Venugopalan, S., & Rohrbach, M. Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1-10 (2016).
51. Gan, Z. et al. Semantic compositional networks for visual captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5630-5639 (2017).
52. Gao, P. et al. A theory of multineuronal dimensionality, dynamics and measurement. Preprint at biorxiv.org/content/biorxiv/early/2017/11/05/214262.full.pdf (2017).
53. Rehurek, R. & Sojka, P. Software framework for topic modelling with large corpora. in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, At Malta 45–50 (2010).
54. McRae, K., Cree, G. S., Seidenberg, M. S. & McNorgan, C. Semantic feature production norms for a large set of living and nonliving things. Behav. Res. Methods, Instruments Comput. **37,** 547–559 (2005).
55. Schakel, A. M. J. & Wilson, B. J. Measuring word significance using distributed representations of words. Preprint at arxiv.org/pdf/1508.02297 (2015).

Acknowledgements

This work was supported in part by the Intel Corporation, the Templeton Foundation, and by NSF REU award #1757554 to T.G.

Author Contributions

M.C.I. and J.D.C. designed the study. M.C.I. and T.G. collected and analyzed the data with input from C.T.E., N.B., and J.D.C. M.C.I., T.G., and J.D.C wrote the manuscript with input from C.T.E. and N.B.

Competing Interests

The authors declare no competing financial interests.