nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-022-01316-8

Semantic projection recovers rich human knowledge of multiple object features from word embeddings

Gabriel Grand[1,2,8], Idan Asher Blank[3,4,8] ✉, Francisco Pereira[5,9] and Evelina Fedorenko[6,7,9]

How is knowledge about word meaning represented in the mental lexicon? Current computational models infer word meanings from lexical co-occurrence patterns. They learn to represent words as vectors in a multidimensional space, wherein words that are used in more similar linguistic contexts—that is, are more semantically related—are located closer together. However, whereas inter-word proximity captures only overall relatedness, human judgements are highly context dependent. For example, dolphins and alligators are similar in size but differ in dangerousness. Here, we use a domain-general method to extract context-dependent relationships from word embeddings: 'semantic projection' of word-vectors onto lines that represent features such as size (the line connecting the words 'small' and 'big') or danger ('safe' to 'dangerous'), analogous to 'mental scales'. This method recovers human judgements across various object categories and properties. Thus, the geometry of word embeddings explicitly represents a wealth of context-dependent world knowledge.

When we say that we know a word, what kind of knowledge do we mean we have? Words allow us to communicate the content of one human mind to another: they are representations of mental structures. More precisely, we use words to express concepts[1]. Concepts correspond to our knowledge about regularities in the world—they are generalizations about the kinds of things that exist and the properties that they have. Words, in turn, associate such abstract knowledge with surface forms of sounds/letters/signs, and differences between word meanings correspond to many key distinctions that we can make between things, properties and events in the world[2–5]. Consequently, the psycholinguistic study of word meanings in the mental lexicon (lexical semantics; see, for example, refs. [6–9]) is necessarily tightly linked to the cognitive study of the architecture of world knowledge in the human mind (semantic memory[10,11]).

However, our world knowledge is broad, detailed and complex. Even an intuitively simple concept such as DOG (here and elsewhere we use all-caps to denote concepts) encompasses rich information about the animal it refers to, including its appearance, biological properties, behavioural tendencies, cultural roles, etc. Only a subset of this conceptual knowledge is communicated by the word 'dog', and different subsets may be communicated in different contexts (for a review, see ref. [12]). Therefore, any theory of lexical semantics should specify the kinds of world knowledge that are captured in the lexicon (for example, ref. [13]). In other words, such theories should identify which subset of semantic memory can be mapped onto language-specific representations of word meaning. Whereas this question is phrased in terms of correspondence between two memory stores (conceptual knowledge and language knowledge), one may alternatively phrase it in terms of a causal relationship between the two, that is, as a question about learnability: What kinds of world knowledge can—in principle—be implicitly learned from the patterns of language use?

Here, we focus on the kinds of world knowledge that might be embedded in a particular pattern of language use: how often different words co-occur with each other. We focus on word co-occurrences for two reasons. First, humans implicitly track such patterns with exquisite accuracy[14], and start doing so early in development[15], so word co-occurrences constitute a core part of our knowledge about language—that is, it is part of the mental lexicon—and it influences linguistic processing[16–21]. Such data are consistent with a hypothesis that dates back to the origins of modern linguistics, namely that word meanings are influenced by their patterns of usage, that is, by the words they tend to appear with (the 'distributional hypothesis')[22–27]. If co-occurrence patterns are part of the mental lexicon, then any kind of world knowledge that is (implicitly) embedded in such patterns and can be recovered from them is in effect stored in the mental lexicon. Second, word co-occurrences are a very simple form of language knowledge. It does not explicitly include, for example, syntactic relationships between words, or word-internal structure. Focusing on word co-occurrence patterns thus allows us to ask which kinds of conceptual knowledge can, in principle, be derived 'bottom up', based on a very simple mechanism of statistical learning over word-forms.

One approach for addressing this question is via computational methods. If a machine is granted access only to word-forms, with no a priori concepts, one can probe the semantic distinctions that are recoverable from the statistics of natural language alone. This approach applies the 'distributional hypothesis' to the artificial minds of machines. Specifically, by tracking the distribution of word co-occurrences in multi-billion word corpora, unsupervised algorithms can learn a representation of word meanings as vectors in a multidimensional space, where the proximity between these vectors increases with the co-occurrence probability of the corresponding words (for a related approach, see refs. [28–30]).

[1]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA. [2]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA, USA. [3]Department of Psychology, UCLA, Los Angeles, CA, USA. [4]Department of Linguistics, UCLA, Los Angeles, CA, USA. [5]National Institute of Mental Health, Bethesda, MD, USA. [6]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, USA. [7]McGovern Institute for Brain Research, MIT, Cambridge, MA, USA. [8]These authors contributed equally: Gabriel Grand and Idan Asher Blank. [9]These authors jointly supervised this work: Francisco Pereira and Evelina Fedorenko. ✉e-mail: iblank@psych.ucla.edu

The resulting space is called a 'word embedding' or a 'distributional semantic model'[31–35].

Recent research has shown that inter-word distances in word embeddings correlate with human ratings of semantic similarity[35–38]. Furthermore, these distances are geometrically consistent across different word pairs that share a common semantic relation. For instance, the location of $\overrightarrow{\text{man}}$ relative to $\overrightarrow{\text{woman}}$ is similar to the location of $\overrightarrow{\text{king}}$ relative to $\overrightarrow{\text{queen}}$. This consistency allows for geometric operations to simulate some conceptual relations, for example, $\overrightarrow{\text{king}} - \overrightarrow{\text{man}} + \overrightarrow{\text{woman}} \approx \overrightarrow{\text{queen}}$[39,40] (but see refs. [41,42]). These results demonstrate that, by simply tracking co-occurrence statistics, a machine with no a priori concepts can obtain a lexicon that contains certain kinds of semantic memory (albeit ungrounded in non-linguistic—for example, perceptual, motor and/or emotional—experience).

Despite these impressive capabilities, word embeddings appear to have a fundamental limitation: the proximity between any two word-vectors captures only a single, semantically rigid measure of overall similarity. In contrast, humans evaluate the conceptual similarity between items in semantic memory flexibly, in a context-dependent manner. Consider, for example, our knowledge of DOLPHINS and ALLIGATORS: when we compare the two on a mental scale of size, from small to big, they are relatively similar; in terms of their intelligence—on a scale from stupid to smart—they are somewhat different; and in terms of danger to us—on a scale from safe to dangerous—they differ significantly. Can such distinct relationships be inferred from word co-occurrence statistics? If so, how is such complex knowledge represented in word embeddings?

Here, we suggest that such conceptual knowledge is present in the structure of word embeddings and use a powerful, domain-general solution for extracting it: 'semantic projection' of word-vectors onto 'feature subspaces' that represent different features (or, more generally, contexts). For instance, to recover the similarities in size amongst nouns in a certain category (for example, animals), we project their representations onto the line that extends from the word-vector $\overrightarrow{\text{small}}$ to the word-vector $\overrightarrow{\text{big}}$ (a 'semantic differential'[43,44]); to compare their levels of intelligence, we project them onto the line connecting $\overrightarrow{\text{stupid}}$ and $\overrightarrow{\text{smart}}$; and to order them according to how dangerous they are, we project them onto the line connecting $\overrightarrow{\text{safe}}$ and $\overrightarrow{\text{dangerous}}$ (for an animation of this procedure, see Supplementary Video 1). We demonstrate that the resulting feature-wise similarities robustly predict human judgements across a wide range of everyday object categories and semantic features. These results corroborate evidence that rich conceptual knowledge can be extracted bottom up from the statistics of natural language, and establish that it is explicitly represented in the geometry of word embeddings, which can be flexibly manipulated in a simple, elegant manner to recover it. Thus, such rich conceptual knowledge must be stored in the mental lexicon.

Results

The rationale of semantic projection. Semantic projection is a domain-general method for comparing word-vectors in the context of a certain semantic feature. A guiding example for applying this method in a simplified, three-dimensional word embedding space (for illustrative purposes) is depicted in Fig. 1 for the category 'animals' and the feature 'size'. Intuitively, to compare animals in terms of this feature, we construct a scale—that is, a straight line in the word embedding space—on which animals could be ordered according to their size (Fig. 1, red line). This scale is constructed via a simple heuristic. We draw a line between antonyms—for example, the word-vector $\overrightarrow{\text{small}}$ and the word-vector $\overrightarrow{\text{large}}$—that denote opposite values of the feature 'size'[43,44] (Fig. 1, red circles). This heuristic corresponds to taking a vector difference: "size" $= \overrightarrow{\text{large}} - \overrightarrow{\text{small}}$ (we use double quotation marks to distinguish between our scale, obtained by subtracting two word-vectors, and the vector of the lexical entry $\overrightarrow{\text{size}}$). Then, by projecting word-vectors of different animals onto this scale, we can approximate common knowledge about their relative sizes. For example, to estimate the relative size of a horse, we would compute the inner product $\overrightarrow{\text{horse}} \cdot$ "size" (in Fig. 1, this orthogonal projection is represented by the blue line extending from the blue dot of $\overrightarrow{\text{horse}}$ to the red scale of "size").

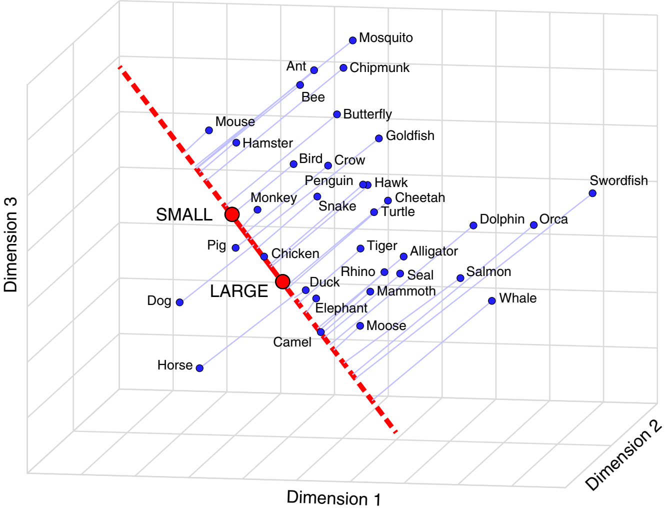


Fig. 1 | Schematic illustration of semantic projection. Word-vectors in the category 'animals' (blue circles) are orthogonally projected (light-blue lines) onto the feature subspace for 'size' (red line), defined as the vector difference between $\overrightarrow{\text{large}}$ and $\overrightarrow{\text{small}}$ (red circles). The three dimensions in this figure are arbitrary and were chosen via principal component analysis to enhance visualization (the original GloVe word embedding has 300 dimensions, and projection happens in that space). For an animated version of this figure, see Supplementary Video 1.

The scale thus created is a one-dimensional subspace in which the feature 'size' governs similarity patterns between word-vectors such that, for example, $\overrightarrow{\text{horse}}$ and $\overrightarrow{\text{tiger}}$ are located close to each other because they are relatively similar in size (in Fig. 1 these two word-vectors, denoted by blue circles, map onto nearby locations on the red line denoting the scale). Critically, these size-related similarity patterns might be different from the global similarities in the original space where, for example, $\overrightarrow{\text{horse}}$ and $\overrightarrow{\text{tiger}}$ might be farther apart: despite their similarity in size (and other features), horses are perceived to be much less dangerous than tigers, belong to a different taxonomic order, occupy different habitats, etc. (in the schematic illustration of Fig. 1, the blue circle corresponding to $\overrightarrow{\text{horse}}$ is relatively far from that of $\overrightarrow{\text{tiger}}$, which happens to be closer to $\overrightarrow{\text{rhino}}$ and $\overrightarrow{\text{alligator}}$; see also Fig. 2).

Note that 'size' is a semantic feature that applies to numerous categories of objects: not only animals, but also mythological creatures, world cities, states of the United States, etc. For each such category, its members could be projected onto the same size scale described above. Hence, semantic projection on a 'feature subspace' is a domain-general method. In this study, we limit ourselves to semantic features that can be represented by one-dimensional subspaces ('scales'). However, other feature subspaces for other semantic features could be of higher dimensionality (Discussion).

Predicting human ratings using semantic projection. We tested whether semantic projection could recover context-dependent conceptual knowledge. To operationalize context-dependent knowledge, we tested how objects from a given semantic category were rated based on a particular semantic feature. Overall, we ran 52 experiments, each testing a different category–feature pair. Pairs

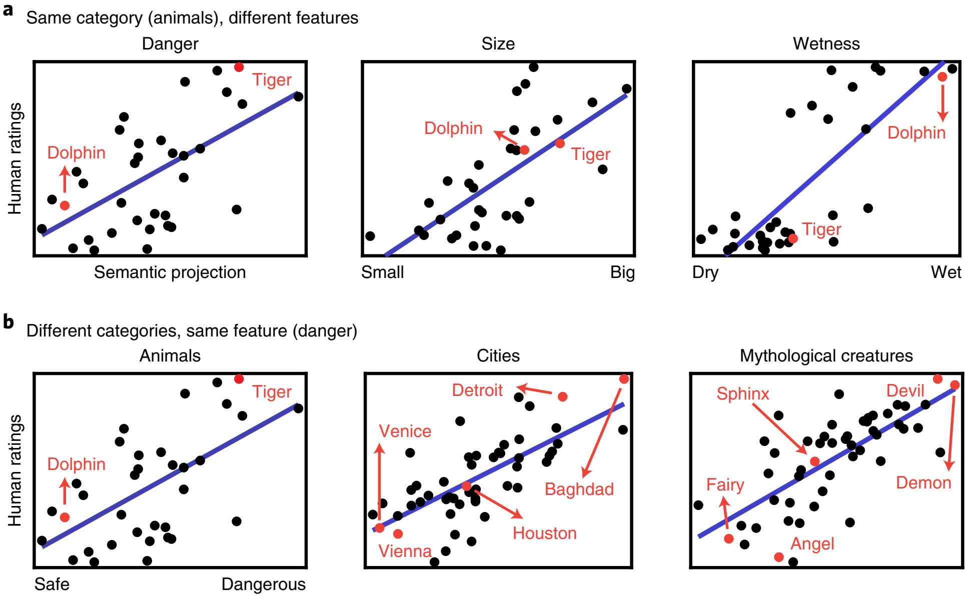


Fig. 2 | Semantic projection predicts human judgements: sample cases. a, Examples of three features for the same category (animals). Notice that the items—for instance, dolphin versus tiger—change their similarities to one another depending on context (feature), and semantic projection recovers these cross-feature differences. In other words, the model does not recover the same relationships across features. **b**, Examples of three categories for the same feature (danger). Sample items are highlighted in red for illustrative purposes. For descriptive and inferential statistics, see Table 1. Each panel is based on data from $n = 25$ participants.

were distributed over nine object categories: animals, clothing, professions, weather phenomena, sports, mythological creatures, world cities, states of the United States and first names (most categories consisted of 50 items). Each category was matched with a subset of the following 17 semantic features: age, arousal, cost, danger, gender (here, limited to a man–woman continuum), intelligence, location (indoors versus outdoors), partisanship (liberal versus conservative), religiosity, size, speed, temperature, valence, (auditory) volume, wealth, weight and wetness. The matching of particular categories to particular features (to create the 52 tested pairs) relied on a combination of an online behavioural norming study and intuitive judgements by the authors (Methods).

To implement semantic projection, we chose GloVe as the word embedding[37] (for results obtained with other models—FastText, word2vec, ELMo and BERT—see Supplementary Results and Extended Data Figs. 1–9). For each test case, we used linear projection to order category items along a line in GloVe that represented a particular semantic feature. This line (that is, one-dimensional 'feature subspace') was computed from vector differences between several antonym pairs that represent opposite ends of the feature continuum. For instance, the continuum for the size feature was anchored by the antonyms $\left\{\overrightarrow{\text{large}}, \overrightarrow{\text{big}}, \overrightarrow{\text{huge}}\right\}$ and $\left\{\overrightarrow{\text{small}}, \overrightarrow{\text{little}}, \overrightarrow{\text{tiny}}\right\}$; the subspace $\overrightarrow{/\!/\text{size}/\!/}$ was the average of $3 \times 3 = 9$ pairwise lines between these antonyms. Onto this subspace, we projected the word-vectors of all category items.

Then, we quantified whether the result corresponded to behavioural ratings collected online from human participants ($n = 25$ per experiment, for a total of 1,400 participants). Figure 2 shows scatterplots of human ratings against model predictions for some illustrative cases. Figures 3 and 4 and Table 1, show the results for each category–feature pair. The correspondence between semantic projection and human data was evaluated using two measures: (1) Pearson's moment correlation and (2) 'pairwise order consistency' (OC_p), that is, the percentage of item pairs (i, j), out of all possible pairings, whose ordering was consistent between humans and semantic projection (for example, for the feature 'size', both rated i as 'bigger' than j). The former is a strict measure of linear relationship, but potentially biased by outliers. The latter is a more lenient measure of correlation, but is more robust to outliers.

Overall, semantic projection successfully recovered human semantic knowledge. Moderate to strong correlations in ratings ($r > 0.5$) were observed for nearly half of the pairs (25/52), and across all experiments, the median correlation was 0.47 (95% CI 0.38–0.55, IQR 0.21–0.65). Similarly, across all experiments, the median OC_p was 65% (95% CI 62–68%, IQR 58–74%). We also adjusted these statistics based on split-half reliability of the behavioural ratings, which is an estimate of the noise ceiling, or 'upper bound', for our measurements (median split-half reliability across experiments: $r = 0.94$, $OC_p = 88\%$; Methods). The 'adjusted median correlation' was 0.52 (that is, the variability in human ratings captured by semantic projection was 27% of the upper bound, $\sqrt{0.27} = 0.52$) (95% CI 0.41–0.60, IQR 0.23–0.74); the 'adjusted median OC_p' increased to 74% (that is, the pairwise order consistency in human ratings captured by semantic projection was 74% of the upper bound) (95% CI 72–78%, IQR 67–86%). In about half of the experiments (32/52), both the correlation and OC_p measures were significant, as evaluated by permutation tests and corrected for multiple comparisons (Methods). Figure 5 summarizes the distribution of the evaluation measures across all experiments, both before and after normalizing these measures relative to split-half reliability. However, the fit between semantic projection and human ratings varied a lot across experiments, with the best fit observed for ratings of names by gender ($r = 0.94$, $OC_p = 87\%$) and the worst fits observed for ratings of cities by cost ($r = -0.15$, $OC_p = 47\%$) and professions by location ($r = -0.12$, $OC_p = 45\%$). We note that, across the 52 experiments, our evaluation measures did not correlate strongly with their corresponding split-half reliability measures (that is, upper bounds), so the cases where semantic projection fails to recover human knowledge are not simply those cases where humans disagree with one another in their judgements. Specifically, whereas the correlation between (Fisher-transformed) r and the split-half reliability of

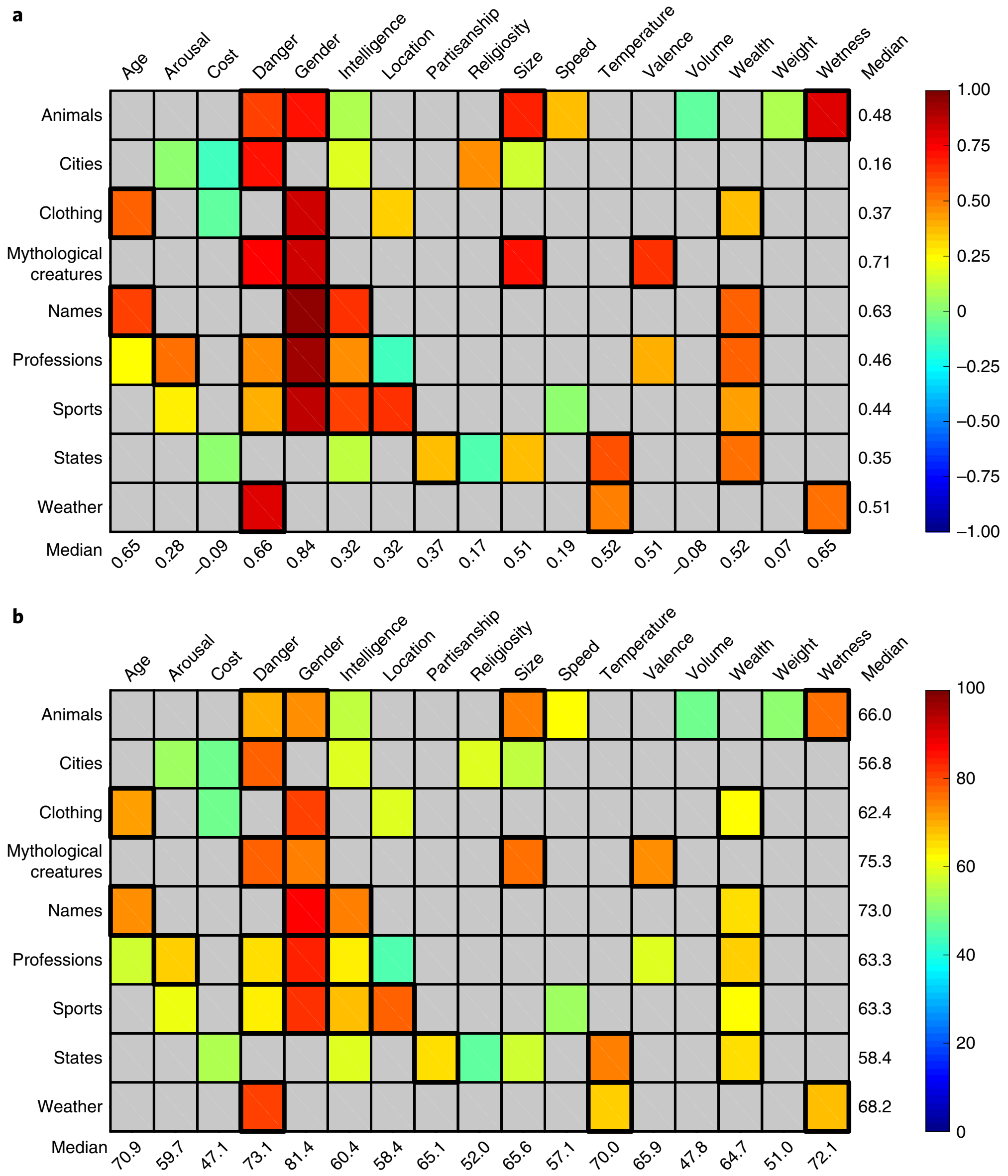


Fig. 3 | Semantic projection predicts human judgements across diverse categories and features. In each panel, each entry in the matrix has a colour corresponding to the strength of fit between semantic projection and human judgements for a given pairing of category (row) and feature (column). Statistically significant fits are surrounded with a thicker contour. For each category–feature pair, ratings were collected from $n = 25$ participants; pairs that were not tested in our experiments are coloured in grey. **a**,**b**, Evaluation measures: Pearson's correlation (**a**, median across experiments: 0.47, 95% CI 0.38–0.55) and pairwise order consistency (**b**, OC_p; median across experiments: 65%, 95% CI 62–68%). In each panel, median values per category are shown on the right whilst median values per feature are shown at the bottom. Evaluation measures are not adjusted for split-half reliability of behavioural ratings.

behavioural ratings across these experiments was 0.39 ($t_{(50)} = 2.96$, $P = 0.002$, based on 10,000 random permutations of experiment labels; 95% CI 0.09–0.61, based on 10,000 bootstrap samples), the correlation between OC_p and the split-half reliability of behavioural ratings was 0.17 ($t_{(50)} = 1.19$, $P = 0.12$, 95% CI −0.06 to 0.40). We further consider this across-experiment variability in the Discussion.

Both ends of a scale contribute to semantic projection. In defining a feature subspace (for example, $//\overrightarrow{\text{size}}//$) for semantic projection, we relied on antonymous adjectives (for example, $\overrightarrow{\text{small}}$ and $\overrightarrow{\text{big}}$) that can be intuitively thought of as defining opposite ends of a scale. The arithmetic difference between the antonymous word-vectors situates this scale within the word embedding space; that is, it defines a 'diagnostic' direction in the space along which concrete objects show high variation with regard to the feature in question. However, this subtraction might not be necessary if the diagnostic information were already sufficiently represented by each adjective vector on its own. In other words, the word-vector $\overrightarrow{\text{big}}$ might already define a size-related diagnostic direction in the space, such that subtracting $\overrightarrow{\text{small}}$ from it is redundant (this would be possible if, for example, these antonyms lay on opposite sides of the origin in the space). To test whether one end of a feature scale may be sufficient for predicting human knowledge, we repeated all experiments but this time compared behavioural ratings with semantic projection on either end of a feature scale (for example, for 'size', projection was performed once on the average of $\left\{\overrightarrow{\text{large}}, \overrightarrow{\text{big}}, \overrightarrow{\text{huge}}\right\}$ and, separately, on the average of $\left\{\overrightarrow{\text{small}}, \overrightarrow{\text{little}}, \overrightarrow{\text{tiny}}\right\}$).

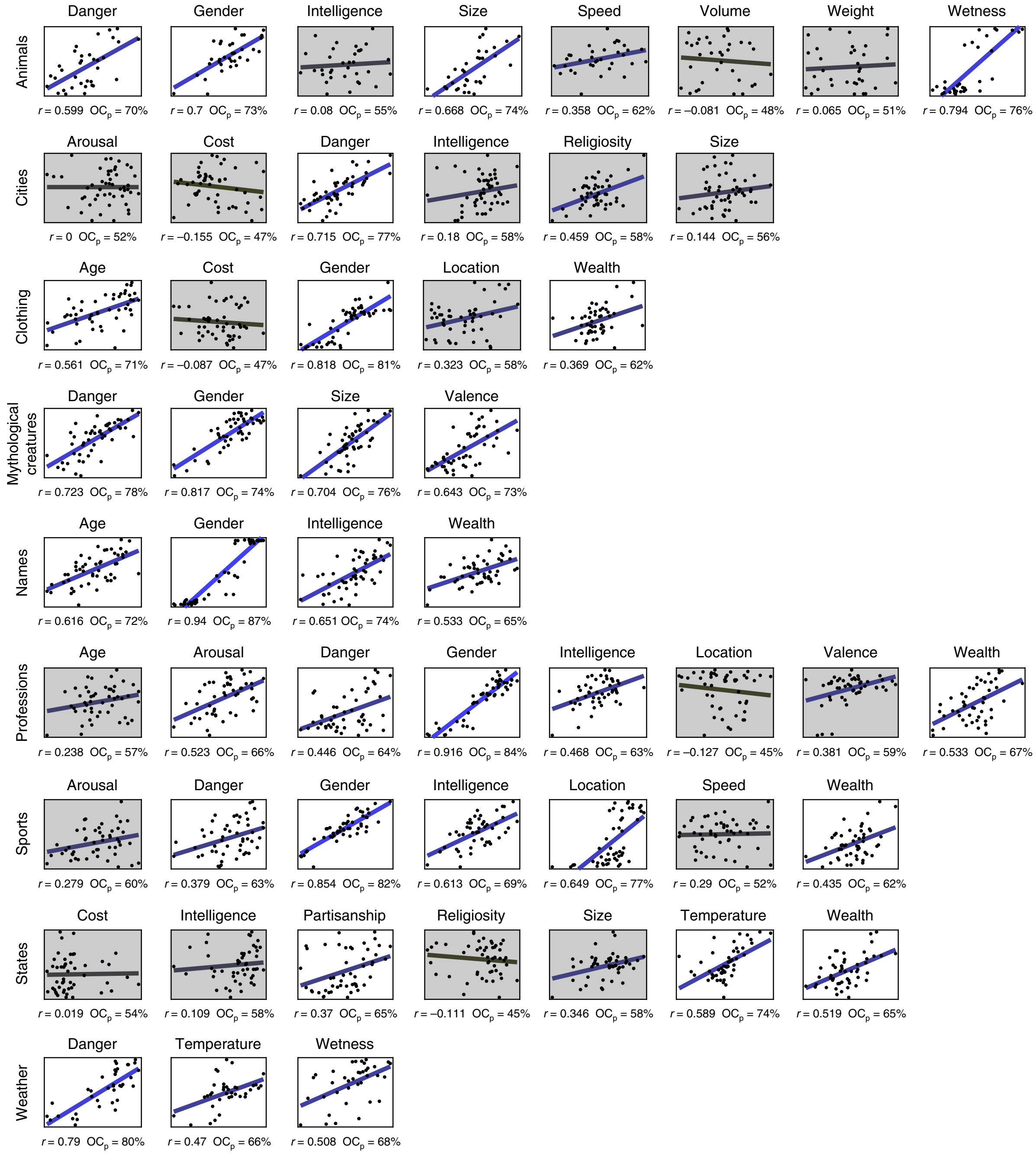


Fig. 4 | Semantic projection predicts human judgements: detailed results. For each of 52 category–feature pairs, scatterplots show the relationship between z scores of average item ratings across participants ($n=25$; y axis) and ratings predicted by semantic projection (x axis). Correlation and pairwise order consistency (OC_p) values are presented below each scatterplot. Experiments for which both of these measures were significant are shown over a white background. Straight lines are linear regression fits to the data and, across figures, vary according to correlation strength from black (weak) to blue (strong).

Across the 52 experiments, projections on a single end had a median correlation of $r=0.06$ with human ratings. This projection scheme performed worse than our original semantic projection even when, for each category–feature pair, we chose from amongst the two ends of a feature scale the one that resulted in a better fit to human judgements ($t_{(51)}=5.30$, $P<0.001$, $d=0.74$, 95% CI 0.20–0.45, one-tailed test). Similar patterns emerged for the other evaluation score, namely pairwise order consistency. Across experiments, the alternative projection scheme had a median value of $\mathrm{OC}_\mathrm{p}=53\%$, worse than our original semantic projection ($t_{(51)}=4.04$, $P<0.001$, $d=0.57$, 95% CI 0.04–0.11). Performance was even worse when, instead of using projection, we computed the distance between each item in a category and either end of the feature scale, using either cosine or Euclidean distance. Therefore, the difference between the vectors of antonymous adjectives—rather than either vector in isolation—provides feature-specific 'diagnostic' dimensions in the word embedding space (see also ref. [45]).

Semantic projection is successful even without outlier items. Might the correlation between human ratings and semantic projection have resulted from a few outliers that were rated as having extreme feature values by both humans and our method? For instance, when rating ani-

Table 1 | Semantic projection predicts human judgements: detailed results*

Category	Feature	Pearson's *r*				OC_p			
		r	*z*	*P* (FDR)	95% CI	OC_p	*z*	*P* (FDR)	95% CI
Clothing	Age	0.56	4.34	<0.001	0.35 to 0.74	0.71	4.28	<0.001	0.61 to 0.76
Names	Age	0.62	4.95	<0.001	0.44 to 0.76	0.72	4.59	<0.001	0.63 to 0.77
Professions	Age	0.24	1.65	0.33	−0.08 to 0.51	0.57	1.38	0.52	0.45 to 0.65
Cities	Arousal	0	−0.002	1	−0.26 to 0.29	0.52	0.46	1	0.41 to 0.61
Professions	Arousal	0.52	3.92	<0.001	0.32 to 0.69	0.66	3.20	0.008	0.56 to 0.72
Sports	Arousal	0.28	2.00	0.15	−0.01 to 0.55	0.60	2.02	0.16	0.48 to 0.69
Cities	Cost	−0.16	−1.07	1	−0.46 to 0.17	0.47	−0.61	1	0.36 to 0.57
Clothing	Cost	−0.09	−0.61	1	−0.32 to 0.17	0.47	−0.60	1	0.37 to 0.55
States	Cost	0.02	0.12	1	−0.20 to 0.27	0.54	0.91	1	0.44 to 0.62
Animals	Danger	0.60	3.89	<0.001	0.35 to 0.76	0.70	3.30	0.006	0.57 to 0.76
Cities	Danger	0.71	6.12	<0.001	0.57 to 0.82	0.77	5.45	<0.001	0.68 to 0.81
Mythological creatures	Danger	0.72	6.30	<0.001	0.58 to 0.84	0.78	5.78	<0.001	0.69 to 0.83
Professions	Danger	0.45	3.23	0.005	0.17 to 0.67	0.64	2.83	0.02	0.53 to 0.72
Sports	Danger	0.38	2.70	0.028	0.18 to 0.55	0.63	2.68	0.03	0.54 to 0.69
Weather	Danger	0.79	6.28	<0.001	0.67 to 0.88	0.80	5.19	<0.001	0.70 to 0.83
Animals	Gender	0.70	4.83	<0.001	0.45 to 0.85	0.73	3.83	<0.001	0.60 to 0.80
Clothing	Gender	0.81	7.83	<0.001	0.73 to 0.88	0.81	6.34	<0.001	0.74 to 0.84
Mythological creatures	Gender	0.82	7.78	<0.001	0.68 to 0.89	0.74	4.93	<0.001	0.65 to 0.79
Names	Gender	0.94	11.87	<0.001	0.90 to 0.97	0.87	7.62	<0.001	0.81 to 0.89
Professions	Gender	0.92	10.64	<0.001	0.87 to 0.95	0.84	6.90	<0.001	0.76 to 0.87
Sports	Gender	0.85	8.67	<0.001	0.76 to 0.92	0.82	6.50	<0.001	0.73 to 0.86
Animals	Intelligence	0.08	0.44	1	−0.26 to 0.42	0.55	0.77	1	0.40 to 0.66
Cities	Intelligence	0.18	1.24	0.674	−0.14 to 0.51	0.58	1.60	0.36	0.45 to 0.67
Names	Intelligence	0.65	5.32	<0.001	0.48 to 0.79	0.74	4.86	<0.001	0.65 to 0.79
Professions	Intelligence	0.47	3.45	0.003	0.21 to 0.67	0.62	2.54	0.045	0.53 to 0.69
Sports	Intelligence	0.61	4.92	<0.001	0.42 to 0.75	0.69	3.86	<0.001	0.58 to 0.75
States	Intelligence	0.11	0.75	1	−0.15 to 0.37	0.58	1.71	0.31	0.48 to 0.67
Clothing	Location	0.32	2.27	0.084	0.01 to 0.57	0.58	1.70	0.30	0.46 to 0.68
Professions	Location	−0.13	−0.86	1	−0.37 to 0.11	0.45	−1.04	1	0.35 to 0.53
Sports	Location	0.65	5.27	<0.001	0.52 to 0.76	0.77	5.50	<0.001	0.68 to 0.82
Animals	Loudness	−0.08	−0.45	1	−0.43 to 0.28	0.48	−0.37	1	0.34 to 0.59
States	Political	0.37	2.67	0.029	0.10 to 0.62	0.65	3.09	0.010	0.54 to 0.73
Cities	Religiosity	0.46	3.43	0.003	0.06 to 0.70	0.58	1.74	0.300	0.46 to 0.68
States	Religiosity	−0.11	−0.76	1	−0.35 to 0.15	0.45	−0.93	1	0.35 to 0.55
Animals	Size	0.67	4.48	<0.001	0.51 to 0.80	0.74	3.94	<0.001	0.63 to 0.78
Cities	Size	0.14	0.99	0.99	−0.10 to 0.37	0.56	1.19	0.708	0.45 to 0.64
Mythological creatures	Size	0.70	5.94	<0.001	0.54 to 0.82	0.76	5.33	<0.001	0.67 to 0.81
States	Size	0.35	2.43	0.055	0.05 to 0.57	0.58	1.53	0.41	0.46 to 0.66
Animals	Speed	0.36	2.10	0.124	0.12 to 0.58	0.62	2.06	0.14	0.50 to 0.70
Sports	Speed	0.03	0.20	1	−0.29 to 0.35	0.52	0.38	1	0.41 to 0.61
States	Temperature	0.57	4.42	<0.001	0.32 to 0.79	0.74	4.88	<0.001	0.63 to 0.81
Weather	Temperature	0.47	2.96	0.013	0.18 to 0.68	0.66	2.79	0.024	0.53 to 0.74
Mythological creatures	Valence	0.64	5.20	<0.001	0.50 to 0.77	0.73	4.73	<0.001	0.65 to 0.78
Professions	Valence	0.38	2.72	0.027	0.12 to 0.58	0.59	1.73	0.298	0.48 to 0.66
Clothing	Wealth	0.37	2.68	0.029	0.04 to 0.65	0.62	2.57	0.044	0.51 to 0.71
Names	Wealth	0.53	4.07	<0.001	0.29 to 0.70	0.65	3.19	0.008	0.55 to 0.72
Professions	Wealth	0.53	4.00	<0.001	0.32 to 0.70	0.67	3.37	0.005	0.56 to 0.74
Sports	Wealth	0.43	3.25	0.005	0.14 to 0.64	0.62	2.54	0.046	0.51 to 0.70
States	Wealth	0.52	3.91	<0.001	0.30 to 0.67	0.65	2.98	0.014	0.54 to 0.72
Animals	Weight	0.07	0.37	1	−0.32 to 0.42	0.51	0.17	1	0.37 to 0.63
Animals	Wetness	0.79	6.01	<0.001	0.66 to 0.89	0.76	4.34	<0.001	0.65 to 0.81
Weather	Wetness	0.51	3.27	0.005	0.22 to 0.73	0.68	3.17	0.008	0.56 to 0.76

*Pearson's correlations are 'raw' values (not Fisher transformed); *z* scores are computed relative to the mean and s.d. of an empirical null distribution (based on 10,000 permutations), estimated with a Gaussian fit (one-tailed test); *P* values across the 52 experiments are corrected for multiple comparisons using FDR; Confidence intervals are empirically estimated based on 10,000 bootstrap re-samples of the data.

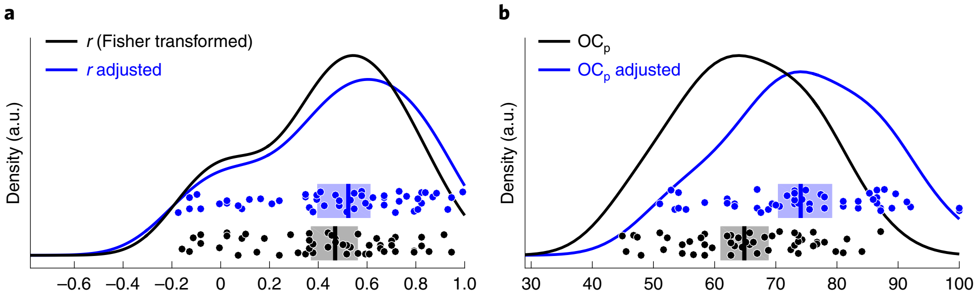


Fig. 5 | Distribution of evaluation scores for semantic projection. a, Pearson's correlation. **b**, Pairwise order consistency. Histograms across 52 category–feature pairs for linear correlations (left) and pairwise order consistency (OC$_p$; right) between semantic projection ratings and human judgements ($n=25$ participants per category–feature pair), before (black) and after (blue) adjustment for inter-rater reliability in ratings. Points show category–feature pairs, vertical lines show medians and semi-transparent rectangles show 95% confidence intervals. Curves show approximated data distributions, with density (in arbitrary units) on the *y* axis.

mals by size, semantic projection might be able to predict that whales are very big whilst mice are very small based on word co-occurrence statistics: the words 'big', 'large' or 'huge' (which define one end of the size subspace) might co-occur with some frequency around 'whale', whereas the words 'small', 'little' or 'tiny' (which define the other end of the size subspace) might co-occur with some frequency around 'mouse'. Such extreme items could cause a strong fit between semantic projection and human ratings even if, for most animals with less extreme values, semantic projection cannot recover human knowledge. However, we do not believe this to be the case, because semantic projection showed high pairwise order consistency with human ratings, and this measure is not strongly biased by outliers.

Nevertheless, to address this concern more directly, we repeated our analyses for each of the 32 significant experiments after removing the two items that received the most extreme (highest and lowest) average ratings. Then, we removed the next two most extreme items, repeated the analyses, and continued this process until 10 items (that is, 20% of the 50-item categories) had been removed. The results of this analysis are presented in Fig. 6. As extreme items were gradually removed from the dataset, both evaluation measures decreased somewhat. Across experiments, the median initial correlation of 0.61 decreased to 0.47 after removing the ten most extreme items. Similarly, the median initial OC$_p$ of 73% decreased to 66%. However, removing extreme items also resulted in less reliable human ratings (reflecting relatively higher noise or uncertainty for less extreme items). Across experiments, the median split-half reliability across behavioural ratings decreased from $r=0.95$ to $r=0.90$, and from OC$_p$ = 89% to OC$_p$ = 85%.

To test whether our evaluation measures were more severely affected by the removal of extreme items, compared with the inter-participant reliability measures, we carried out a mixed-effects linear regression (using the R package lme4 (ref. [46])) to predict the values of both measures based on: (1) the number of removed items, (2) the type of measure (raw versus upper bound/reliability) and (3) the interaction between (1) and (2). Random intercepts, and random slopes for (1), were included by category and by feature. For both evaluation measures, the interaction term did not significantly improve the model (modelling *r* and its split-half reliability: $\chi^2_{(1)}=2.14$, $P=0.14$, $b=-0.02$, 95% CI −0.04 to 0.005; modelling OC$_p$ and its split-half reliability: $\chi^2_{(1)}=0.001$, $P=0.97$, $b<-10^{-4}$, 95% CI −0.004 to 0.003). In other words, the decrease in the evaluation measures was statistically indistinguishable from the decrease in the corresponding split-half reliability across participants. We therefore do not find evidence that extreme feature values bias the success of semantic projection beyond the extent to which they increase certainty, or reduce noise, in semantic knowledge itself (as reflected by increased reliability of human judgements).

Discussion

We find that semantic projection of concrete and/or imageable nouns onto feature subspaces (that is, 'semantic differentials'[43,44]) can approximate human ratings of the corresponding entities along multiple, distinct feature continua. The method we use is simple yet robust and broadly applicable, successfully predicting human judgements across a range of everyday object categories and semantic features. These results demonstrate that semantic knowledge about context-dependent similarities is explicitly represented in the structure of word embeddings. Thus, extremely detailed conceptual knowledge can be constructed bottom up by merely tracking word co-occurrence statistics, and it can be expressed in a simple representational system (cf. deep neural networks where context sensitivity is engineered into the system via, for example, recurrence[47] or attention[48,49] mechanisms). These findings are consistent with previous reports demonstrating the richness of knowledge in word embeddings[50,51], including knowledge of social biases[52–55], and expert knowledge in specialized domains including, perhaps, yet-to-be-made scientific discoveries[56].

We go beyond these prior studies in three key respects. First, we cover a wide range of conceptual categories and their properties, including abstract and/or infrequently named properties. Second, where previous studies have relied on supervised learning algorithms, which are trained directly on human data to learn a mapping of word-vectors onto specific features[40,57–61], semantic projection is nearly unsupervised. It only requires that the task be specified in terms of antonyms (or, more broadly, 'context-defining' anchors), but can then generate predictions without any access to human data (except, of course, to evaluate its outputs). In this sense, our model does not require any training. The choice of antonyms amounts to the same level of supervision that is introduced in any evaluation that probes the quality of fixed word embeddings by choosing particular words and simply evaluating the similarity between their respective representations. Third, whereas supervised learning algorithms can learn any linear combination of the dimensions in a word embedding in order to optimize their prediction of human data, semantic projection guarantees a solution that is interpretable. It a priori defines a line that is intuitively analogous to a 'scale'. With semantic projection, a single, general-purpose word embedding (unbiased towards any particular domain; cf. ref. [62]) can be re-structured to emphasize different properties. Therefore, its semantic knowledge is not only detailed, but can also be flexibly used. This is a hallmark of the human cognitive architecture, because language use is never 'context-less'. Any reasonable model of the mental lexicon should employ context-sensitive representations.

The nature of semantic knowledge representations has been long debated, and continues to be central to the study of the human mind

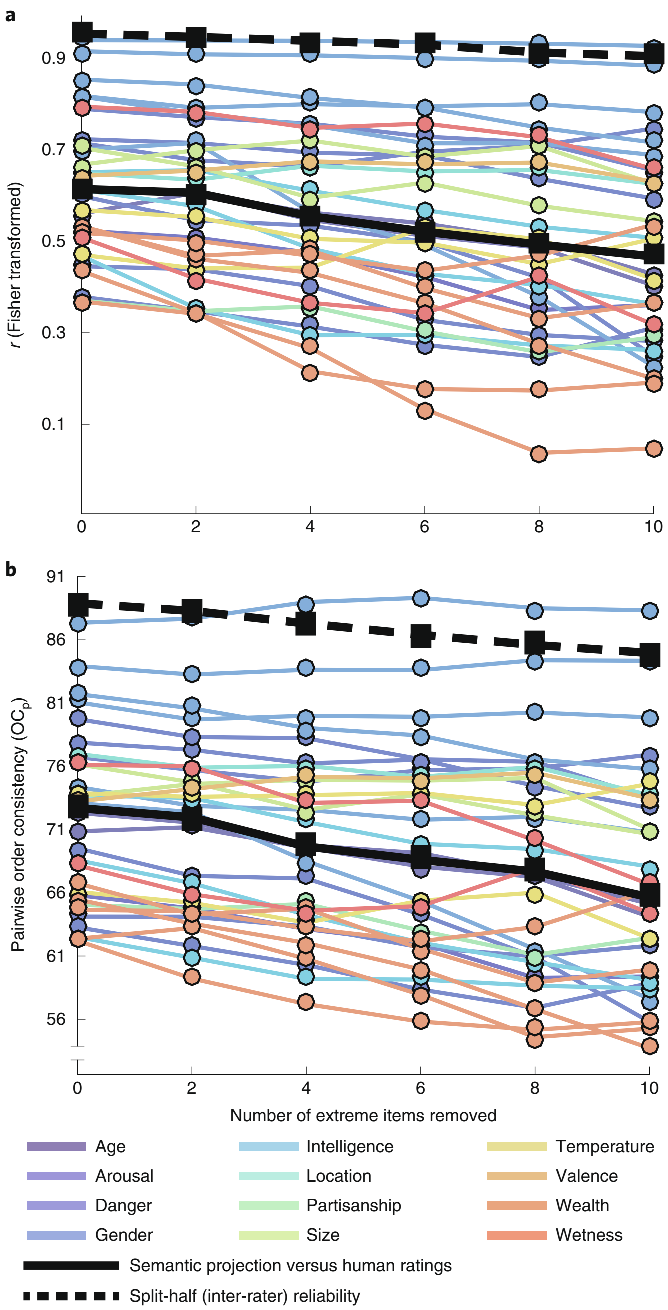


Fig. 6 | The fit of semantic projection to human ratings is not driven by outliers. a,**b**, r (**a**, Fisher transformed) and OC$_p$ values (**b**) plotted as functions of the number of items with the most extreme human ratings that were removed from the data (note that, for **b**, the y axis is 'broken'). Data from each of 32 category–feature pairs (those with significant results in our main experiment) are shown, with colour varying by feature. Continuous black lines show the average across the 32 pairs. Dotted black lines show the upper bound (split-half reliability across participants) averaged across these pairs. Notice that the continuous and black lines are mostly parallel: as extreme items are removed, the ability of semantic projection to recover human knowledge decreases, but the reliability of human knowledge itself similarly decreases. When predicting the decrease in these measures, the interaction between the number of removed items and the type of measure (critical measure versus its upper bound) is not significant (modelling r and its split-half reliability: $\chi^2_{(1)} = 2.14$, $P = 0.14$, $b = -0.02$, 95% CI −0.04 to 0.005; modelling OC$_p$ and its split-half reliability: $\chi^2_{(1)} = 0.001$, $P = 0.97$, $b < -10^{-4}$, 95% CI −0.004 to 0.003). Human data for each category/feature pair are based on $n = 25$ participants.

in cognitive science, neuroscience and philosophy (for reviews, see refs. [4,63–66]). In particular, knowledge about the relationships between object categories and the features characterizing them has been traditionally discussed in the context of symbolic representations such as feature lists[67,68], structured schemata[69] or highly elaborate intuitive theories[70,71] (for an early example of modelling context-dependent similarities, see ref. [72]). Prior studies that have attempted to extract such knowledge from natural corpora have had to augment the tracking of word co-occurrences with more elaborate information such as dependency parses or supervised identification of particular linguistic patterns[73–78]. Our results instead suggest that the distributional, sub-symbolic representational format of word embeddings can support the flexible re-structuring of object categories to highlight specific features. Moreover, given that these spaces are hypothesized to approximate lexical semantics, the current study suggests that complex feature-specific knowledge is part of a word's meaning.

We emphasize that the lexical knowledge in word embeddings is detached from, rather than grounded in, non-linguistic (for example, perceptual, motor or emotional) experience. Therefore, even though the geometry of word embeddings captures human knowledge that, for example, an 'elephant' is 'bigger' than a 'mouse' (a perceptual piece of knowledge), this model does not have access to the perceptual correlates of the words 'elephant', 'big' or 'mouse'. In fact, it only knows the meaning of these words in terms of their relatedness to other words (that is, their intra-linguistic meaning). Nonetheless, our findings provide a proof of principle that world knowledge can be independently acquired from statistical regularities in natural language itself[79].

Therefore, the current study is consistent with the intriguing hypothesis that, like word embedding spaces, humans can use language as a gateway to acquiring conceptual knowledge[50,80–84]. Indeed, humans are sensitive to patterns of word co-occurrence, and use them during language processing[14–21] (for tracking of relationships between words and linguistic contexts more generally, see refs. [85–92]). In addition, evidence from congenitally blind individuals suggests that such patterns may indeed be sufficient for acquiring some forms of perceptual knowledge, for example, similarities amongst colours or actions involving motion, and subtle distinctions between sight-verbs such as 'look', 'see' and 'glance'[93–98]. Thus, in the absence of direct (for example, perceptual) experience, language itself can serve as a source of semantic knowledge. Yet even without conjecturing about learnability, that is, a causal relationship from word co-occurrence patterns to world knowledge, our work supports a more general hypothesis: given that word co-occurrence patterns are stored in the mental lexicon (that is, constitute part of our knowledge about language) and that they contain the kinds of conceptual knowledge studied here, we conclude that the human mental lexicon contains a subset of semantic memory that is more rich, detailed and complex than has traditionally been assumed.

What kinds of category–feature relations could be acquired in this manner? Past research has suggested that word embedding spaces (prior to any semantic projection) capture gross knowledge about the sensory modalities associated with different objects[99], but they fare relatively poorly in approximating detailed perceptual properties in comparison with abstract (for example, encyclopaedic or functional) knowledge[76,100–103]. The current results suggest a more nuanced view (Fig. 4). First, knowledge about some perceptual features (for example, size) was successfully predicted for some categories (animals, mythological creatures; see also ref. [60]). Second, whereas some abstract features (for example, gender and danger) could be recovered via semantic projection across all the categories with which they were paired, other abstract features (for example, intelligence) were only recovered for some categories but not others.

In addition, by re-structuring word embeddings, semantic projection may be able to resolve current debates concerning the particular object properties that are learnable via word co-occurrences (for example, refs. [104–108]). It is possible that some aspects of semantic knowledge may be captured to different degrees in the embedding space, depending on a variety of factors. For instance, sensory modalities may differ from one another in how consistently they are linguistically coded across speakers, and such differences across modalities also vary across languages[109]. The more variable the words for a certain concept are, the fewer contexts of usage are available for each of those words in a corpus, potentially leading to both noisier vector representations for those words and a more diffuse influence of those words as contexts for learning other vectors. As a result, domains of knowledge whose linguistic coding is less consistent across speakers may be less learnable from word embeddings. This is a promising direction of research, in that such effects can be quantified and tested. Furthermore, even if concepts relevant to a given modality are consistently coded across individuals, it might be the case that this modality is not referred to in written language with the same frequency as spoken or signed language, or at all. If so, it would have less influence in structuring the embedding space or not be learnable at all. More generally, the factors underlying the variability in the performance of semantic projection across different feature–category pairs (and across different entities within a category) would be a fruitful area of future investigation.

Notwithstanding its variability in performance, we emphasize that semantic projection exhibits promising generalizability across both categories and features. Our categories span animate and inanimate categories (for example, animals versus clothing), natural and human-made categories (for example, weather versus clothing) and common and proper nouns (for example, professions versus cities). Future work may further test how well this method can recover knowledge about entities that are abstract rather than concrete, as well as about concepts that correspond to parts of speech other than nouns (for example, projecting verbs on subspaces defined by adverb antonym pairs). Similarly, our method generalizes across different 'kinds' of features: from those that are judged to be relatively binary (for example, the wetness of animals) to those that vary more continuously between two extremes (for example, the gender associated with articles of clothing). It may further extend to other kinds of features such as those with multiple, discrete values (for example, 'colour' or 'number of legs') and, more generally, to complex types of context-dependent knowledge represented in semantic subspaces with more than one dimension.

These prospects for generalizability raise deeper questions about semantic knowledge representation in word embeddings: Which sets of word-vectors constitute psychologically plausible categories?[4] Which semantic subspaces represent features (or, more generally, contexts[110,111])? Which categories can be meaningfully described with which features?[112,113] And what geometric operations besides linear projection could capture different kinds of human knowledge? Addressing these questions could help characterize the structure of word embedding spaces and, critically, inform general theories of categories and features that are fundamental to the study of concepts. Specifically, if word embeddings are found to represent information that approximates human patterns of category formation, feature elicitation and context-dependent semantic judgements, then their structure could perhaps provide a principled way for deriving an ontology of concepts.

In conclusion, semantic projection in word embeddings is a powerful method for estimating human knowledge about the structure of categories under distinct contexts. Within the distributional semantics literature, this method continues the tradition of applying simple linear algebraic operations to perform useful semantic comparisons in word embeddings (for example, vector subtraction, cosine similarities and matrix multiplication[114,115]). Moreover, compared with prior attempts at extracting semantic knowledge from patterns of natural language use, this method requires significantly less human supervision and/or corpus annotation. Most importantly, it obviates the need to define a priori a constrained ontology of semantic features that would span the vector space. Given an existing word embedding, which was not constructed based on any such ontology, semantic projection can flexibly recover a variety of semantic features. Therefore, we believe that we have only scratched the surface of the total volume of knowledge captured by word embeddings. We hope that semantic projection will provide a useful, generalizable framework for deeper exploration of such models.

Methods

All experimental procedures were approved by Massachusetts Institute of Technology's Committee on the Use of Humans as Experimental Subjects. Informed consent was obtained from all participants, as required by the committee.

Materials. We created a set of categories and features that met five criteria: (1) each category consisted of concrete and/or imageable objects/entities and had at least 30 members (most had 50); (2) each category could be characterized with respect to several features (to evaluate whether semantic projection generalized across features for a given category); (3) each feature was applicable to several categories (to evaluate whether semantic projection generalized across categories for a given feature); (4) each feature was one-dimensional; (5) categories and features spanned diverse aspects of everyday semantic knowledge. All categories and their constituent items were selected from an extensive set of nouns, generated for a large-scale study on lexical memory[116].

Concept categories. Our materials included nine semantic categories: animals, clothing, professions, weather phenomena, sports, mythological creatures, world cities, states of the United States and first names. These categories have been used in feature elicitation studies, with varying degrees of prevalence[111,117–121]. The first four have been used frequently, and the next two less. The last three have been used rarely and, unlike the other categories, consisted of proper nouns. Items in all categories were used as cues in a mega-study of lexical knowledge using word associations[122] (http://w3.usf.edu/FreeAssociation/).

Most categories consisted of 50 representative items (for example, the clothing category included words such as 'hat', 'tuxedo', 'sandals' and 'skirt'; see Supplementary Methods for the full set of materials). The only exceptions were the categories of animals (34 items), weather phenomena (37 items) and professions (49 items). To ensure that items were representative of their respective category, we chose from a superset of nouns within each category (from ref. [116]) the 50 most frequent nouns according to the SubtlexUS Word Frequency Database[123]. We discarded multi-word expressions (except for states of the United States, for example, 'North Dakota') and words that did not appear in the vocabulary of the word embedding (see below).

For two categories, additional selection criteria were used to increase the variability for the 'gender' feature: for the first names category, we chose the 20 most common male names, the 20 most common female names and the 10 most common unisex names from the past 100 years, based on public data from the US Social Security Administration (www.ssa.gov/oact/babynames/decades/century.html; github.com/fivethirtyeight/data/tree/master/unisex-names). For the 'professions' category, efforts were made to balance the items by stereotypical gender, because the majority of the most frequently occurring profession nouns were traditionally male.

Semantic features. Our materials included 17 semantic features, each associated with a subset of the categories above: size, temperature, valence, (auditory) volume, speed, location (indoors versus outdoors), intelligence, wetness, weight, wealth, gender (here, limited to a man–woman continuum), danger, age, religiosity, partisanship (liberal versus conservative), cost and arousal. These features have been produced in feature elicitation studies, with varying degrees of prevalence[111,119,120]: the first four almost invariably, the next three frequently, the next six less frequently and the last four very rarely.

Choosing appropriate features for each concept category. For each category in our set of materials, only some features provided meaningful contexts for rating category members. For instance, sports could be rated by 'danger' but not by 'size', and names could be rated by 'gender' but not by 'temperature'. To select category–feature pairings to be used in the experiment, we relied on a combination of our own intuitive judgements and ratings obtained in a norming study. In particular, out of the set of $9 \times 17 = 153$ possible category–feature pairings, we first selected a subset of 45 pairs that appeared intuitively appropriate. In parallel, we asked participants ($n = 50$) on Amazon Mechanical Turk[124] (MTurk, www.mturk.com) to rate how likely they were to describe each category (for example, animals) in terms of each feature (for example, size). Instead of using feature names, we used the antonyms that represented opposite extreme values of that feature. For instance, a typical question read, 'How likely are you to describe animals as large/big/huge or small/little/tiny?'. The 153 category–feature pairs were rated on a scale from 1 ('not likely at all') to 5 ('extremely likely').

We averaged the ratings of each category–feature pair across participants and selected those for which the mean rating exceeded the 75th percentile (corresponding to a mean rating of 3.44 or above on the likelihood scale). This subset consisted of 39 pairs. We further removed two pairs where the norming question was ambiguous or otherwise unclear: animals–age (we are likely to describe an animal as being young or old, but comparing different species in terms of their age is less meaningful) and cities–partisanship (we are likely to describe US cities as being liberal or conservative, but many items in this category were cities in other countries).

The remaining 37 pairs partially overlapped with the manually selected subset (26 pairs in common). The union of these two subsets, totalling 56 pairs, was used in the main experiment (four pairs were later removed because human knowledge about them was extremely noisy; see below). The inclusion of manually selected pairs that were not rated highly in the norming study could only weaken the performance of semantic projection. If human knowledge about these 'unlikely' category–feature pairs is noisy or uncertain, we do not expect to recover it from a word embedding via semantic projection. Despite there being other category–feature pairings that could be deemed appropriate for testing, we limited our study to 56 experiments because this number offered breadth of coverage whilst still being manageable given our resources.

Computational model. *The GloVe word embedding.* We chose to conduct our experiment in the GloVe word embedding[37], because it outperforms several other word embeddings in predicting word similarity judgements[38]. We used 300-dimensional GloVe vectors derived from the Common Crawl corpus (http://commoncrawl.org/), which contains approximately 42 billion uncased tokens and a total vocabulary size of 1.9 million. To limit the vocabulary to words with robust co-occurrence estimates, we considered only the 500,000 most frequent words. See Supplementary Methods and Extended Data Figs. 1–9 for results obtained with two other word embeddings, as well as two deep language models that are trained to represent words in context (that is, process sentences).

Defining feature subspaces. A one-dimensional feature subspace is approximated by the vector difference between antonyms that represent opposite ends of the feature continuum. In our implementation, each end was represented by three words similar in meaning that were chosen by the authors based on intuition from amongst the words used as cues in the word association study in ref. [122]. For instance, for the feature 'danger', one end was represented by the three vectors $\{\overrightarrow{\text{dangerous}}, \overrightarrow{\text{deadly}}, \overrightarrow{\text{threatening}}\}$, and the other end by the three vectors $\{\overrightarrow{\text{safe}}, \overrightarrow{\text{harmless}}, \overrightarrow{\text{calm}}\}$. The feature subspace was then defined as the average of the $3\times3=9$ possible vector differences (or 'lines') between the two ends.

This averaging procedure was used to obtain more robust approximations of feature subspaces that were not strongly dependent on (1) the particular choice of antonyms by the authors or (2) the representation of each individual antonym resulting from the particular training regime and corpus used to generate the GloVe space. Indeed, lines for a given feature subspace were not always strongly aligned, with the mean cosine similarity between one line and the average of the remaining eight being 0.533 (corresponding to an angle of 0.99π or 57°). Nevertheless, lines were still more aligned within a feature subspace than across subspaces: the mean cosine similarity between one line and the average of lines from another subspace was 0.095, which suggests that the subspaces are effectively orthogonal to each other.

Human ratings. Common knowledge about the 56 category–feature pairs in our dataset was evaluated on MTurk. Each category–feature pair was evaluated in a separate ~5-min-long experiment with $n=25$ participants (for a total of 1,400 participants; age and gender information were not collected). Participants were paid US $4 for rating the items in a single category according to a single feature. Each item had a separate sliding scale from 0 to 100, where 0 corresponded to one end of the feature continuum (for example, 'small, little or tiny' for size) and 100 corresponded to the other end (for example, 'large, big or huge'). The words describing each end of the scale were the same words that were used to define the feature subspaces in the GloVe space.

We limited participation to MTurk users in the United States who had previously completed at least 1,000 experiments ('human information tasks' (HITs)) with an acceptance rate of 95% or above. To account for participant idiosyncrasies in how the scale was used, we z-scored the ratings of each participant. To exclude participants who had provided random responses, we computed Pearson's moment correlation coefficient between the scaled responses of each participant and the average of the scaled responses across the rest of the sample. For each experiment, this procedure thus resulted in 25 inter-subject correlation (ISC) values. These ISCs were Fisher transformed to improve the normality of their distribution[125], and participants whose ISC was inferior to the mean ISC in their sample by more than 2.5 s.d. (that is, participants whose ratings showed weak correlations with the rest of the group) were removed from further analysis. In the majority of experiments, no participants were excluded, and no more than two participants were excluded from any given experiment.

For a given category–feature pair, the average ISC across participants provides a measure of the noise, or uncertainty, in common knowledge about that pair. When examining the average ISC for each experiment, we identified four outlier experiments for which ISCs were low (<0.07): cities by temperature, cities by wealth, clothing by arousal and clothing by size. The ISC values for these four experiments were clear outliers relative to the distribution of ISC values across the remaining 52 experiments and, accordingly, we removed them from further analysis.

We note that the choice of $n=25$ participants per category–feature pair was based on our prior experience with rating studies. In the current dataset, the ISC between a single participant and the remaining 24 participants is nearly the same as the ISC between that participant and a random subset of only 14 other participants (namely, this difference never exceeds $r=0.025$). Thus, the agreement across 25 participants in their ratings is similar to the agreement that would be obtained with 15 participants, demonstrating that our sample size is sufficient for obtaining stable ratings.

Predicting human ratings using semantic projection. For each category–feature pair, we evaluated how well the ratings produced by semantic projection predicted the human ratings, by using two complementary measures as described below. For each measure and each pair, we computed a 95% confidence interval based on 10,000 bootstrap samples. Across category–feature pairs, we computed the median, the inter-quartile range and 95% confidence intervals of the median.

Measure 1: linear correlation. We used Pearson's moment correlation coefficient to estimate how much of the variance in human ratings across items could be explained by the ratings from the semantic projection. This measure is sensitive to even minor shifts in ratings. A slight change in a single item would (in most cases) affect the s.d. of the entire rating distribution and, consequently, change the respective contribution of each item to the correlation value. Therefore, it provides a strict test for semantic projection. Nonetheless, it is strongly biased by outliers. A strong correlation might reflect not the overall quality of semantic projection but, rather, a few extreme ratings made by both humans and the semantic projection method.

Measure 2: pairwise order consistency. This measure, which we denote $\mathrm{OC_p}$, estimated the percentage of item pairs, out of all possible pairings, for which the difference in ratings had the same sign in both human judgements and the semantic projection. For example, in the animals–danger experiment, for every two animals (i, j) such that i was rated by humans as more dangerous than j, we tested whether the semantic projection had predicted the same (versus opposite) pattern. This measure is sensitive only to the direction of pairwise differences but not to their magnitude. For instance, in the animals–danger experiment, humans rate alligators to be more dangerous than dolphins, so we require that semantic projection makes the same judgement regardless of how far apart the two animals fall on the feature subspace. Here, a change in the rating of a single item would only affect those pairs that (1) include this item and (2) reversed the direction of their difference as a result of this change. This measure is therefore robust to outliers. We note that this measure is closely related to Kendall's rank correlation coefficient (tau) but has a more intuitive interpretation due to its range of [0%, 100%] (instead of [−1, 1]).

Significance testing. The significance of both evaluation measures for each experiment (that is, category–feature pair) was quantified via a permutation procedure. For each of 10,000 iterations, we randomly shuffled the labels of category items in the feature subspace (but not their labels in the human data) and recomputed our two evaluation measures to obtain their empirical null distributions. The significance of each veridical measure was then computed relative to the mean and s.d. of its null distribution, estimated with a Gaussian fit (one-tailed test). For each evaluation measure, P values across the 52 experiments were corrected for multiple comparisons using false discovery rate (FDR) correction[126].

We chose to rely on a Gaussian fit rather than the 'raw' distribution of permuted values because many permutation tests returned a count of 0 (that is, no permutation produced a better result than the empirical data). When zero probabilities are submitted to FDR correction, they remain at zero instead of increasing. In contrast, the Gaussian fit produces a P value that is numerically higher than absolute zero and is thus a conservative choice that allows for probabilities to 'correct upward'. Moreover, we reasoned that a Gaussian fit would obtain probabilities that better reflect the full null distribution, not just those permutations that happened to be produced by the randomization code. Evidence for the quality of this fit is provided in the Supplementary Information. All the significant measures we report remain significant when P values from the Gaussian fits are replaced with 'raw' P values based on counting permutations (and all of the non-significant measures remain non-significant).

Estimating an upper bound (noise ceiling) for the evaluation measures. Because semantic projection approximates human knowledge, its success is limited by the amount of noise or uncertainty in that knowledge. Specifically, if human ratings for a certain category–feature pair exhibit low inter-rater reliability, then testing whether semantic projection captures human knowledge for this pair makes little sense, given that people disagree with one another in their judgements. Therefore, for each experiment, we compared our first measure—linear correlation—with the (Fisher-transformed) split-half reliability of ratings across participants. Specifically, we divided the squared correlation (that is, the percentage of variance in human ratings explained by semantic projection) by the squared split-half reliability, and took the square root of the result. We followed a similar procedure for computing split-half pairwise order consistency (OC_p) across participants, to obtain an upper bound for our second measure. Here, we divided the OC_p from the semantic projection by split-half reliability. For both measures, values greater than 1 were set to 1. In the main text, 'adjustment for upper bound' refers to this normalization of our evaluation measures relative to inter-rater reliability in behavioural data.

We chose to use split-half reliability rather than inter-subject correlations (that is, the average agreement between each participant and the rest of the group, or a one-versus-all reliability) because the former is usually higher. Therefore, it provides a more conservative noise estimate against which to normalize our evaluation measures. In other words, we chose a reliability measure that would lead to less inflation in our evaluation measures.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All behavioural data and GloVe vectors as reported in the paper are available on the Open Science Framework (https://osf.io/5r2sz/). The full database of GloVe vectors (including many words not used in this study) is available for download from https://nlp.stanford.edu/projects/glove/.

Code availability

Custom MATLAB codes for replicating the analyses are available on the Open Science Framework page referenced above (https://osf.io/5r2sz/). The outputs of these codes include all data visualized in Figs. 2–6.

Received: 25 May 2020; Accepted: 31 January 2022;
Published online: 14 April 2022

References

1. Marr, D. in *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (ed. Marr, D.) 8–38 (MIT Press, 1982).
2. Goldberg, A. E. *Constructions: A Construction Grammar Approach to Argument Structure* (Univ. of Chicago Press, 1995).
3. Jackendoff, R. *Foundation of Language: Brain, Meaning, Grammar, Evolution* (Oxford Univ. Press, 2002).
4. Murphy, G. *The Big Book of Concepts* (MIT Press, 2004).
5. Jackendoff, R. *A User's Guide to Thought and Meaning* (Oxford Univ. Press, 2012).
6. Steinberg, D. D. & Jakobovits, L. A. *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*. (Cambridge Univ. Press, 1971).
7. Richards, M. M. in *Language Development, Vol. 1: Syntax and Semantics* Vol. 1 (ed. S. Kuczaj) 365–396 (Routledge, 1982).
8. Pinker, S. & Levin, B. *Lexical and Conceptual Semantics* (MIT Press, 1991).
9. Pustejovsky, J. *Semantics and the Lexicon* Vol. 49 (Springer, 2012).
10. Quillian, M. R. *Semantic Memory.* PhD thesis, Carnegie Intitute of Technology (1966).
11. Tulving, E. in *Organization of Memory* Vol. 1 (eds Tulving E. & Donaldson W.) 381–403 (Academic, 1972).
12. Gleitman, L. & Papafragou, A. in *The Oxford Handbook of Cognitive Psychology* (ed D. Resiberg) 255–275 (Oxford Univ. Press, 2013).
13. Jackendoff, R. Parts and boundaries. *Cognition* **41**, 9–45 (1991).
14. Smith, N. J. & Levy, R. The effect of word predictability on reading time is logarithmic. *Cognition* **128**, 302–319 (2013).
15. Skarabela, B., Ota, M., O'Connor, R. & Arnon, I. 'Clap your hands' or 'take your hands'? One-year-olds distinguish between frequent and infrequent multiword phrases. *Cognition* **211**, 104612 (2021).
16. Monsalve, I. F., Frank, S. L. & Vigliocco, G. in *Proc. 13th Conference of the European Chapter of the Association for Computational Linguistics*, 398–408 (Association for Computational Linguistics, 2012).
17. Frank, S. & Thompson, R. Early effects of word surprisal on pupil size during reading. In *Proc. 34th Annual Conference of the Cognitive Science Society* Vol. 34 (eds Miyake, N. et al.) 1554–1559 (Cognitive Science Society, 2012).
18. Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P. & Van den Bosch, A. Prediction during natural language comprehension. *Cereb. Cortex* **26**, 2506–2516 (2015).
19. McDonald, S. & Ramscar, M. Testing the distributional hypothesis: the influence of context on judgements of semantic similarity. In *Proc. 23rd Annual Conference of the Cognitive Science Society* https://escholarship.org/uc/item/6959p7b0 (2001).
20. Ellis, N. C. & Simpson-Vlach, R. Formulaic language in native speakers: triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguist. Linguistic Theory* **5**, 61–78 (2009).
21. Louwerse, M. M. Embodied relations are encoded in language. *Psychonomic Bull. Rev.* **15**, 838–844 (2008).
22. De Saussure, F. *Course in General Linguistics* (Columbia Univ. Press, 2011).
23. Wittgenstein, L. *Philosophical Investigations.* §114–115 (Wiley-Blackwell, 2010).
24. Harris, Z. S. Distributional structure. *Word* **10**, 146–162 (1954).
25. Firth, J. R. in *Studies in Linguistic Analysis Special volume of the Philological Society* (ed. Firth, J. R.) 1–31 (Blackwell, 1957).
26. Miller, G. A. & Charles, W. G. Contextual correlates of semantic similarity. *Lang. Cogn. Process.* **6**, 1–28 (1991).
27. Sahlgren, M. The distributional hypothesis. *Ital. J. Disabil. Stud.* **20**, 33–53 (2008).
28. Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **3**, 1137–1155 (2003).
29. Huang, E. H., Socher, R., Manning, C. D. & Ng, A. Y. in *Proc. 50th Annual Meeting of the Association for Computational Linguistics* Vol. 1: Long Papers, 873–882 (Association for Computational Linguistics, 2012).
30. Collobert, R. et al. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12**, 2493–2537 (2011).
31. Lenci, A. Distributional semantics in linguistic and cognitive research. *Ital. J. Ling.* **20**, 1–31 (2008).
32. Erk, K. Vector space models of word meaning and phrase meaning: a survey. *Lang. Linguist. Compass* **6**, 635–653 (2012).
33. Clark, S. in *Handbook of Contemporary Semantics* (eds Lappin S. & Fox C.) 493–522 (Blackwell, 2015).
34. Turney, P. D. & Pantel, P. From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* **37**, 141–188 (2010).
35. Baroni, M., Dinu, G. & Kruszewski, G. in *Proc. 52nd Annual Meeting of the Association for Computational Linguistics* Vol. 1: Long Papers, 238–247 (Association for Computational Linguistics, 2014).
36. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. in *Proc. 26th International Conference on Neural Information Processing Systems*, 3111–3119 (Curran Associates, Inc., 2013).
37. Pennington, J., Socher, R. & Manning, C. in *Proc. 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543 (Association for Computational Linguistics, 2014).
38. 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).
39. Levy, O. & Goldberg, Y. in *Advances in Neural Information Processing Systems*, 2177–2185 (Curran Associates, Inc., 2014).
40. Lu, H., Wu, Y. N. & Holyoak, K. J. Emergence of analogy from relation learning. *Proc. Natl Acad. Sci. U. S. A.* **116**, 4176–4181 (2019).
41. Rogers, A., Drozd, A. & Li, B. in *Proc. 6th Joint Conference on Lexical and Computational Semantics*, 135–148 (Association for Computational Linguistics, 2017).
42. Peterson, J. C., Chen, D. & Griffiths, T. L. Parallelograms revisited: exploring the limitations of vector space models for simple analogies. *Cognition* **205**, 104440 (2020).
43. Osgood, C. E. The nature and measurement of meaning. *Psychol. Bull.* **49**, 197 (1952).
44. Osgood, C. E. Semantic differential technique in the comparative study of cultures. *Am. Anthropol.* **66**, 171–200 (1964).
45. Kozima, H. & Ito, A. Context-sensitive measurement of word distance by adaptive scaling of a semantic space. In *Proc. RANLP-95,* 161–168 (John Benjamins Publishing Company, 1995).
46. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
47. Peters, M. E. et al. in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2227–2237 (Association for Computational Linguistics, 2018).
48. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1: Long and Short Papers, 4171–4186 (Association for Computational Linguistics, 2019).
49. Vaswani, A. et al. Attention is all you need. In *Proc. Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 5998–6008 (Curran Associates, Inc., 2017).
50. Huebner, P. A. & Willits, J. A. Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Front. Psychol.* **9**, 133 (2018).
51. Unger, L. & Fisher, A. V. The emergence of richly organized semantic knowledge from simple statistics: a synthetic review. *Dev. Rev.* **60**, 100949 (2021).

52. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
53. Lewis, M. & Lupyan, G. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nat. Hum. Behav.* **4**, 1021–1028 (2020).
54. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V. & Kalai, A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proc. 30th International Conference on Neural Information Processing Systems (NIPS 2016)*, 4356–4364 (Curran Associates, Inc., 2016).
55. Kozlowski, A. C., Taddy, M. & Evans, J. A. The geometry of culture: analyzing the meanings of class through word embeddings. *Am. Sociol. Rev.* **84**, 905–949 (2019).
56. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
57. Johns, B. T. & Jones, M. N. Perceptual inference through global lexical similarity. *Top. Cog. Sci.* **4**, 103–120 (2012).
58. Herbelot, A. & Vecchi, E. M. in *Proc. 2015 Conference on Empirical Methods in Natural Language Processing*, 22–32 (Association for Computational Linguistics, 2015).
59. Gupta, A., Boleda, G., Baroni, M. & Padó, S. in *Proc. 2015 Conference on Empirical Methods in Natural Language Processing*, 12–21 (Association for Computational Linguistics, 2015).
60. Utsumi, A. Exploring what is encoded in distributional word vectors: a neurobiologically motivated analysis. *Cogn. Sci.* **44**, e12844 (2020).
61. Ichien, N., Lu, H. & Holyoak, K. J. Predicting patterns of similarity among abstract semantic relations. *J. Exp. Psychol. Learning Memory Cogn.* https://doi.org/10.1037/xlm0001010 (2021).
62. Iordan, M. C., Giallanza, T., Ellis, C. T., Beckage, N. M. & Cohen, J. D. Context matters: recovering human semantic structure from machine learning analysis of large-scale text corpora. *Cogn. Sci.* **46**, e13085 (2022).
63. Laurence, S. & Margolis, E. in *Concepts: Core Readings* (eds Laurence, S. & Margolis, E.) 3–81 (MIT Press, 1999).
64. Markman, A. B. *Knowledge Representation* (Lawrence Erlbaum, 2013).
65. Mahon, B. Z. & Hickok, G. Arguments about the nature of concepts: symbols, embodiment, and beyond. *Psychon. Bull. Rev.* **23**, 941–958 (2016).
66. Yee, E., Jones, M. & McRae, K. in *Stevens' Handbook of Experimental Psychology, Memory and Cognitive Processes* Vol. 2 (ed Wixted J.) (Wiley, 2014).
67. Rosch, E. & Mervis, C. B. Family resemblances: studies in the internal structure of categories. *Cogn. Psychol.* **7**, 573–605 (1975).
68. Smith, E. E. & Medin, D. L. *Categories and Concepts* (Harvard Univ. Press, 1981).
69. Rumelhart, D. & Ortony, A. in *Schooling and the Acquisition of Knowledge* (eds Anderson R. C., Spiro R. J., & Montague W. E.) 99–135 (Lawrence Erlbaum, 1977).
70. Gopnik, A., Meltzoff, A. N. & Bryant, P. *Words, Thoughts, and Theories*, Vol. 1 (MIT Press, 1997).
71. Gopnik, A. in *Chomsky and His Critics* (eds Antony L. & Hornstein N.) 238–254 (Blackwell, 2003).
72. Medin, D. L. & Schaffer, M. M. Context theory of classification learning. *Psych. Rev.* **85**, 207 (1978).
73. Poesio, M. & Almuhareb, A. in *Proc. Association for Computational Linguistics SIGLEX Workshop on Deep Lexical Acquisition*, 18–27 (Association for Computational Linguistics, 2005).
74. Barbu, E. in *Proc. ESSLLI Workshop on Distributional Lexical Semantics*, 9–16 (Association for Logic, Language and Information, 2008).
75. Baroni, M. & Lenci, A. in *Proc. Workshop on Geometrical Models of Natural Language Semantics*, 1–8 (Association for Computational Linguistics, 2009).
76. Baroni, M., Murphy, B., Barbu, E. & Poesio, M. Strudel: a corpus-based semantic model based on properties and types. *Cogn. Sci.* **34**, 222–254 (2010).
77. Rubinstein, D., Levi, E., Schwartz, R. & Rappoport, A. in *Proc. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* Vol. 2: Short Papers, 726–730 (Association for Computational Linguistics, 2015).
78. Kelly, C., Devereux, B. & Korhonen, A. Automatic extraction of property norm-like data from large text corpora. *Cogn. Sci.* **38**, 638–682 (2014).
79. Lupyan, G. & Lewis, M. From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Lang. Cogn. Neurosci.* **34**, 1319–1337 (2019).
80. Rumelhart, D. E. in *Metaphor and Thought* (ed. Andrew Ortony) 71–82 (Cambridge Univ. Press, 1979).
81. Landauer, T. K. & Dumais, S. T. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211–240 (1997).
82. Elman, J. L. An alternative view of the mental lexicon. *Trends Cogn. Sci.* **8**, 301–306 (2004).
83. Elman, J. L. On the meaning of words and dinosaur bones: lexical knowledge without a lexicon. *Cogn. Sci.* **33**, 547–582 (2009).
84. Lupyan, G. & Bergen, B. How language programs the mind. *Top. Cogn. Sci.* **8**, 408–424 (2016).
85. Clifton, C., Frazier, L. & Connine, C. Lexical expectations in sentence comprehension. *J. Verbal Learn. Verbal Behav.* **23**, 696–708 (1984).
86. MacDonald, M. C., Pearlmutter, N. J. & Seidenberg, M. S. The lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* **101**, 676–703 (1994).
87. Trueswell, J. C., Tanenhaus, M. K. & Garnsey, S. M. Semantic influences on parsing: use of thematic role information in syntactic ambiguity resolution. *J. Mem. Lang.* **33**, 285–318 (1994).
88. Garnsey, S. M., Pearlmutter, N. J., Myers, E. & Lotocky, M. A. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *J. Mem. Lang.* **37**, 58–93 (1997).
89. Hale, J. in *Proc. Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, 1–8 (Association for Computational Linguistics, 2001).
90. Traxler, M. J., Morris, R. K. & Seely, R. E. Processing subject and object relative clauses: evidence from eye movements. *J. Mem. Lang.* **47**, 69–90 (2002).
91. Gennari, S. P. & MacDonald, M. C. Semantic indeterminacy in object relative clauses. *J. Mem. Lang.* **58**, 161–187 (2008).
92. Levy, R. P. Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177 (2008).
93. Marmor, G. S. Age at onset of blindness and the development of the semantics of color names. *J. Exp. Child Psych.* **25**, 267–278 (1978).
94. Landau, B. & Gleitman, L. R. *Language and Experience: Evidence from the Blind Child,* Vol. 8 (Harvard Univ. Press, 2009).
95. Shepard, R. N. & Cooper, L. A. Representation of colors in the blind, color-blind, and normally sighted. *Psychol. Sci.* **3**, 97–104 (1992).
96. Noppeney, U., Friston, K. J. & Price, C. J. Effects of visual deprivation on the organization of the semantic system. *Brain* **126**, 1620–1627 (2003).
97. Bedny, M., Caramazza, A., Pascual-Leone, A. & Saxe, R. Typical neural representations of action verbs develop without vision. *Cereb. Cortex* **22**, 286–293 (2011).
98. Bedny, M., Koster-Hale, J., Elli, G., Yazzolino, L. & Saxe, R. There's more to "sparkle" than meets the eye: knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition* **189**, 105–115 (2019).
99. Louwerse, M. & Connell, L. A taste of words: linguistic context and perceptual simulation predict the modality of words. *Cogn. Sci.* **35**, 381–398 (2011).
100. Baroni, M. & Lenci, A. Concepts and properties in word spaces. *Ital. J. Ling.* **20**, 55–88 (2008).
101. Andrews, M., Vigliocco, G. & Vinson, D. Integrating experiential and distributional data to learn semantic representations. *Psych. Rev.* **116**, 463–498 (2009).
102. Riordan, B. & Jones, M. N. Redundancy in perceptual and linguistic experience: comparing feature-based and distributional models of semantic representation. *Top. Cogn. Sci.* **3**, 303–345 (2011).
103. Hill, F., Reichart, R. & Korhonen, A. Simlex-999: eEvaluating semantic models with (genuine) similarity estimation. *Comput. Linguist.* **41**, 665–695 (2015).
104. Kim, J. S., Elli, G. V. & Bedny, M. Knowledge of animal appearance among sighted and blind adults. *Proc. Natl Acad. Sci. U. S. A.* **116**, 11213–11222 (2019).
105. Kim, J. S., Elli, G. V. & Bedny, M. Reply to Ostarek et al.: Language, but not co-occurrence statistics, is useful for learning animal appearance. *Proc. Natl Acad. Sci. U. S. A.* **116**, 21974–21975 (2019).
106. Kim, J. S., Elli, G. V. & Bedny, M. Reply to Lewis et al.: Inference is key to learning appearance from language, for humans and distributional semantic models alike. *Proc. Natl Acad. Sci. U. S. A.* **116**, 19239–19240 (2019).
107. Ostarek, M., Van Paridon, J. & Montero-Melis, G. Sighted people's language is not helpful for blind individuals' acquisition of typical animal colors. *Proc. Natl Acad. Sci. U. S. A.* **116**, 21972–21973 (2019).
108. Lewis, M., Zettersten, M. & Lupyan, G. Distributional semantics as a source of visual knowledge. *Proc. Natl Acad. Sci. U. S. A.* **116**, 19237–19238 (2019).
109. Majid, A. et al. Differential coding of perception in the world's languages. *Proc. Natl Acad. Sci. U. S. A.* **115**, 11369–11376 (2018).
110. Clark, E. V. in *Cognitive Development and the Acquisition of Language* (ed. Moskowitz B. A.) 223–260 (Academic, 1973).
111. Binder, J. R. et al. Toward a brain-based componential semantic representation. *Cogn. Neuropsychol.* **33**, 130–174 (2016).
112. Barsalou, L. W. & Sewell, D. R. Contrasting the representation of scripts and categories. *J. Mem. Lang.* **24**, 646–665 (1985).
113. Tanaka, J. W. & Taylor, M. Object categories and expertise: is the basic level in the eye of the beholder? *Cogn. Psych.* **23**, 457–482 (1991).
114. Baroni, M. & Zamparelli, R. in *Proc. 2010 Conference on Empirical Methods in Natural Language Processing*, 1183–1193 (Association for Computational Linguistics, 2010).
115. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. in *NeurIPS (prev. NIPS)*, 3111–3119 (Curran Associates, Inc., 2013).
116. Mahowald, K., Isola, P., Fedorenko, E., Gibson, E. & Oliva, A. Memorable words are monogamous: the role of synonymy and homonymy in word recognition memory. Preprint at *PsyArxiv* https://psyarxiv.com/p6kv9/ (2018).

117. Paivio, A., Yuille, J. C. & Madigan, S. A. Concreteness, imagery, and meaningfulness values for 925 nouns. *J. Exp. Psychol.* **76**, 1–25 (1968).
118. Battig, W. F. & Montague, W. E. Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. *J. Exp. Psychol.* **80**, 1–46 (1969).
119. Cree, G. S. & McRae, K. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *J. Exp. Psychol. Gen.* **132**, 163–201 (2003).
120. 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* **37**, 547–559 (2005).
121. Pereira, F., Botvinick, M. & Detre, G. Using Wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artif. Intell.* **194**, 240–252 (2013).
122. Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. The University of South Florida free association, rhyme, and word fragment norms. *Behav. Res. Methods* **36**, 402–407 (2004).
123. Brysbaert, M. & New, B. Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Beh. Res. Methods* **41**, 977–990 (2009).
124. Buhrmester, M., Kwang, T. & Gosling, S. D. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* **6**, 3–5 (2011).
125. Silver, N. C. & Dunlap, W. P. Averaging correlation coefficients: should Fisher's *z* transformation be used? *J. Appl. Psychol.* **72**, 146 (1987).
126. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).

Acknowledgements

F.P. was supported by the National Institute of Mental Health Intramural Research Program ZIC-MH002968. E.F. was supported by R01 awards DC016607 and DC016950, U01 award NS6945189 and funds from the McGovern Institute for Brain Research, the Department of Brain and Cognitive Sciences and the Simons Center for the Social Brain. This work was partially supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Research Laboratory (AFRL), under contract FA8650-14-C-7358 (to I.A.B., F.P. and E.F.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The authors are responsible for all aspects of the study, and views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL or the US Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon.

Author contributions

G.G. and I.A.B. are co-first authors. F.P. and E.F. are co-senior authors. Conceptualization: G.G. and I.A.B. Methodology (behavioural experiments): G.G., F.P. and E.F. Behavioural data collection: G.G. Data curation: I.A.B. Model implementation: G.G. and I.A.B. Statistical analysis: I.A.B. Visualization: I.A.B. Writing, original draft: I.A.B. Writing, review and editing: G.G., I.A.B., F.P. and E.F. Funding acquisition: F.P. and E.F. Supervision: F.P. and E.F.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-022-01316-8.

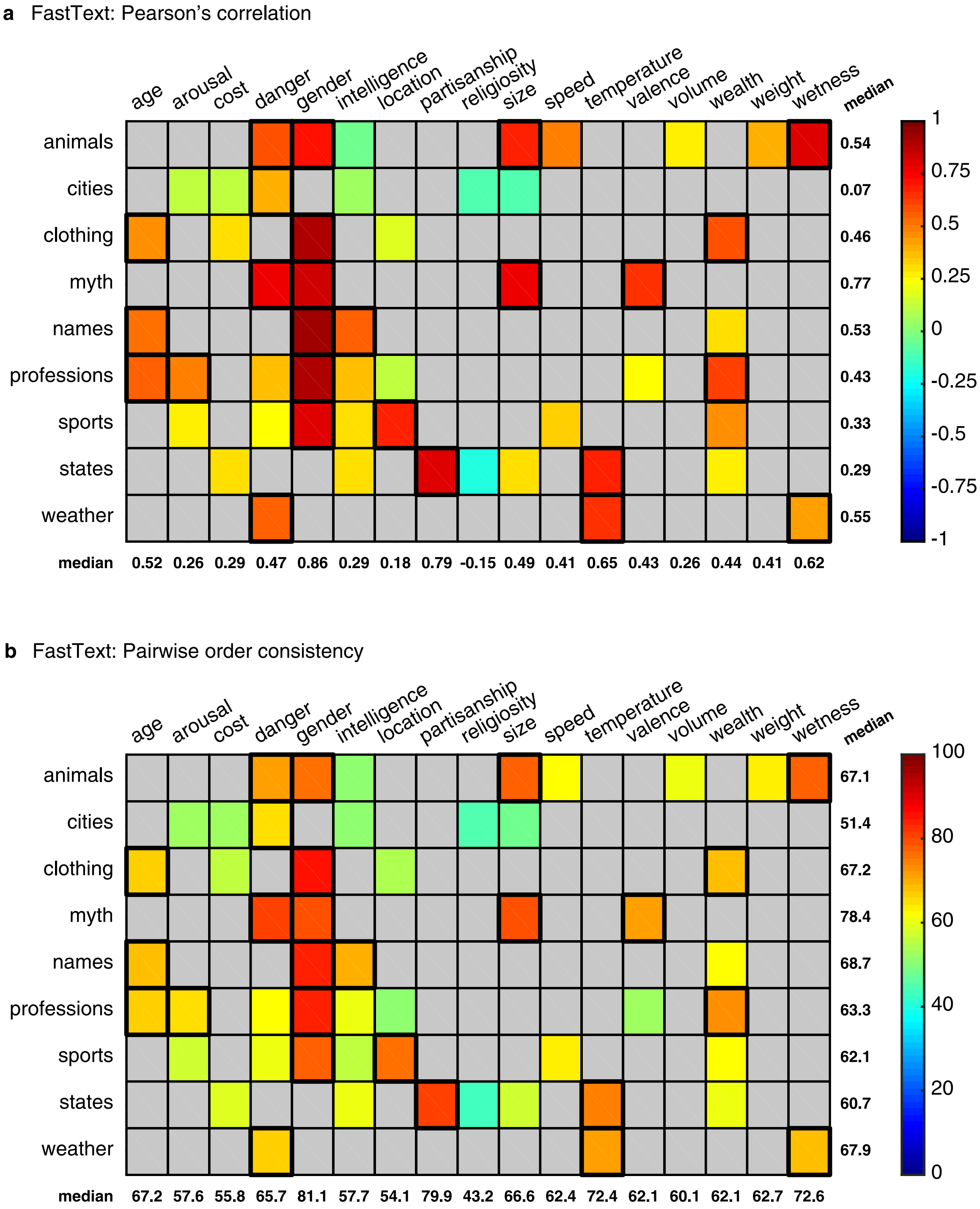
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01316-8.

Correspondence and requests for materials should be addressed to Idan Asher Blank.

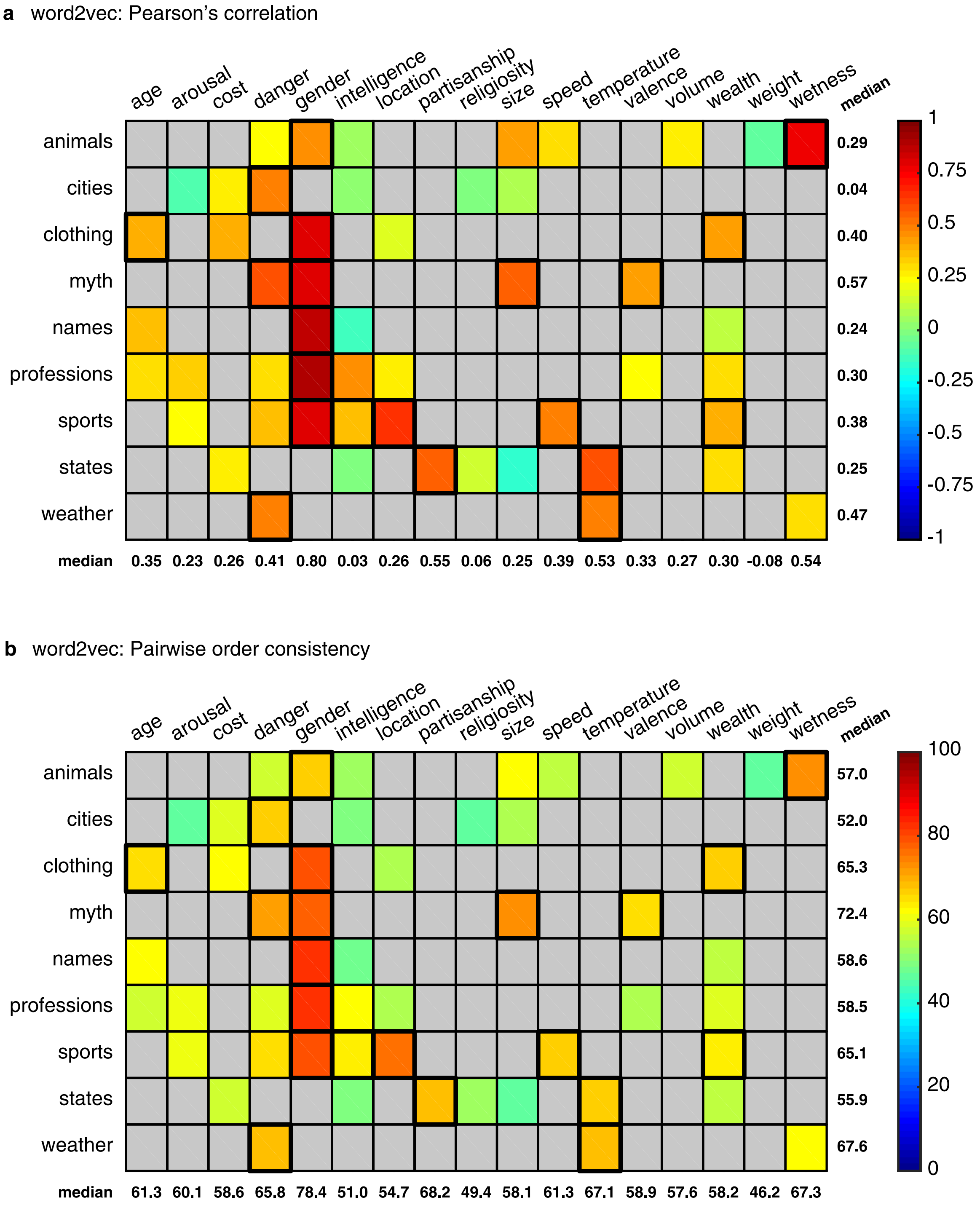
Peer review information *Nature Human Behaviour* thanks Gary Lupyan, Katrin Erk and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

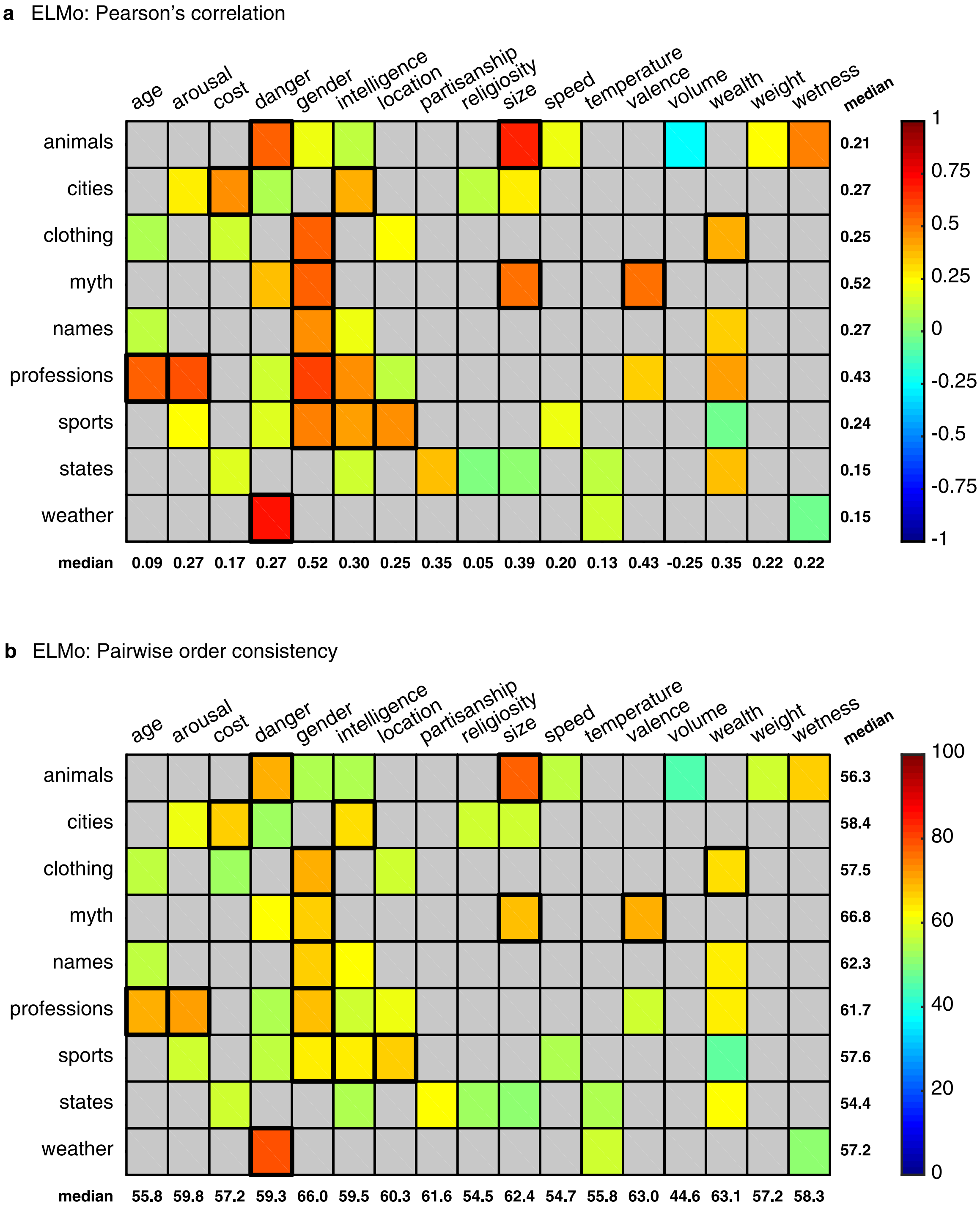
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Correspondence between human judgments and semantic projection using FastText. Conventions are the same as in Fig. 3 in the manuscript. Descriptive statistics across all tested pairs: (a) Pearson's correlation: med = 0.41 (CI_{95} = 0.29-0.50, IQR = 0.26-0.57), adjusted med = 0.44 (CI_{95} = 0.35-0.53, IQR = 0.29-0.60). (b) OC_p: med = 64% (CI_{95} = 61-68%, IQR = 57-73%), adjusted med = 73% (CI_{95} = 70-78%, IQR = 67-81%).

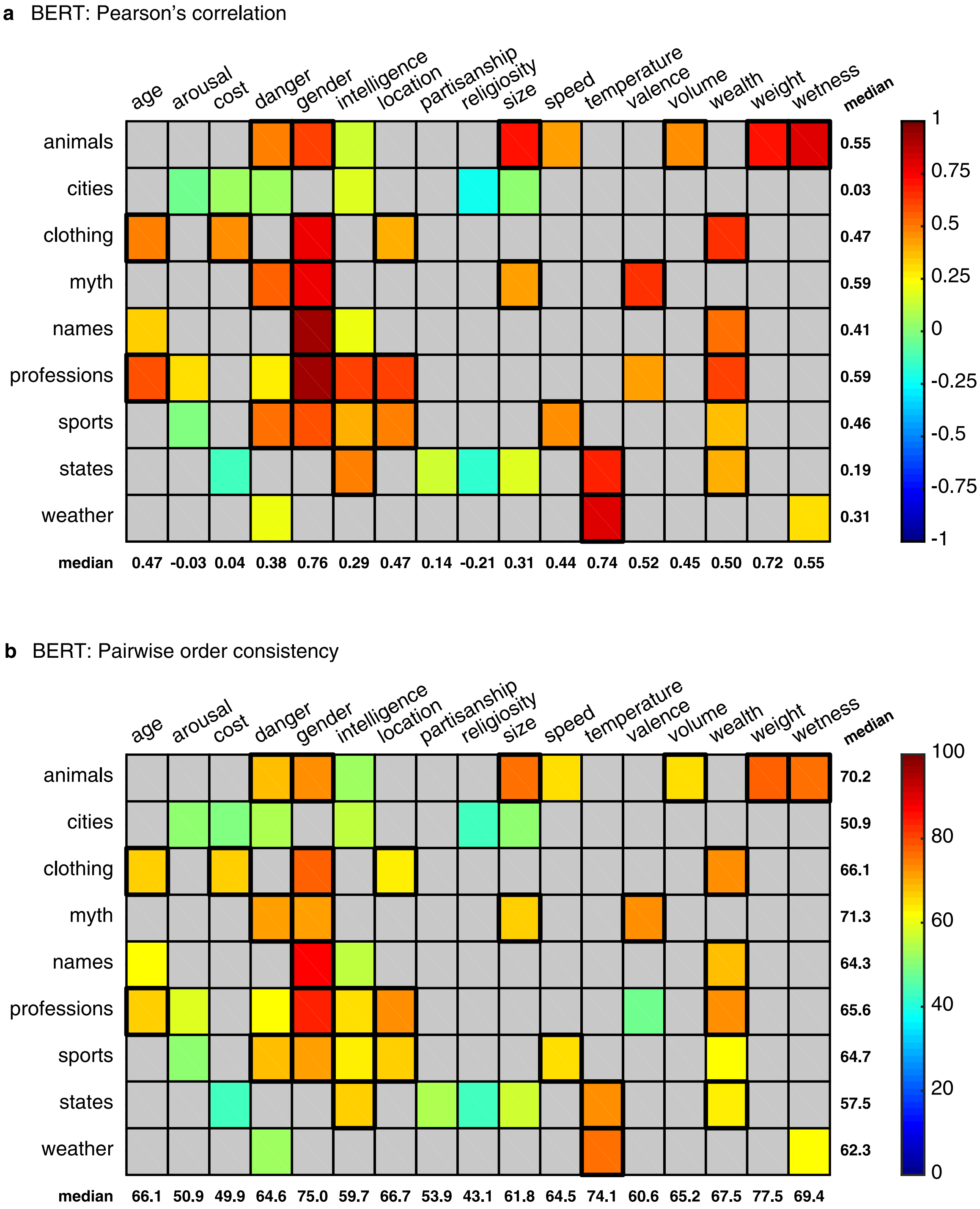


Extended Data Fig. 2 | Correspondence between human judgments and semantic projection using word2vec. Conventions are the same as in Fig. 3 in the manuscript. Descriptive statistics across all tested pairs: (a) Pearson's correlation: med = 0.33 (CI_{95} = 0.27-0.40, IQR = 0.22-0.44), adjusted med = 0.35 (CI_{95} = 0.28-0.43, IQR = 0.24-0.47). (b) OC_p: med = 62% (CI_{95} = 57-65%, IQR = 55-67%), adjusted med = 71% (CI_{95} = 65-74%, IQR = 63-78%).

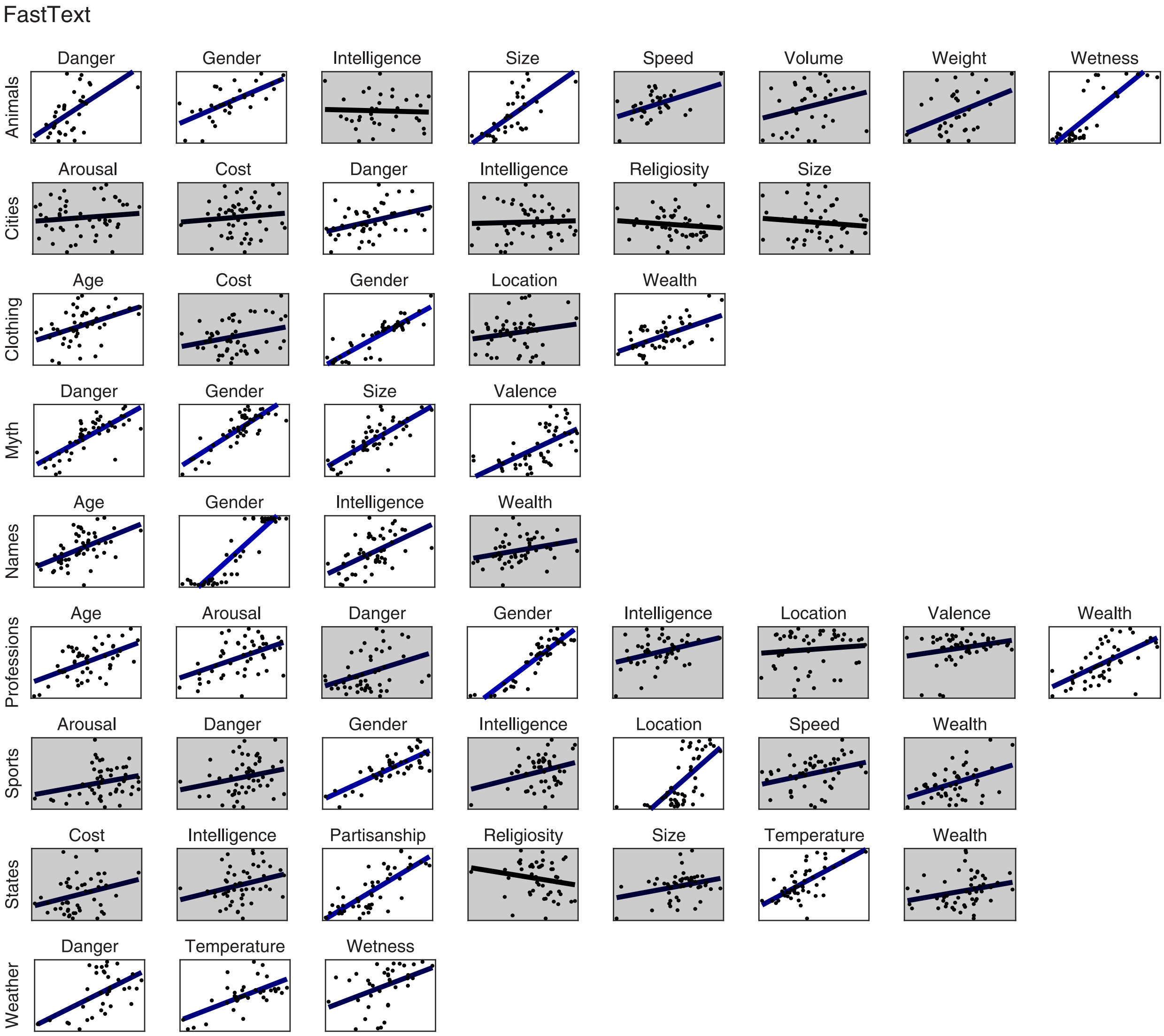

a ELMo: Pearson's correlation

b ELMo: Pairwise order consistency

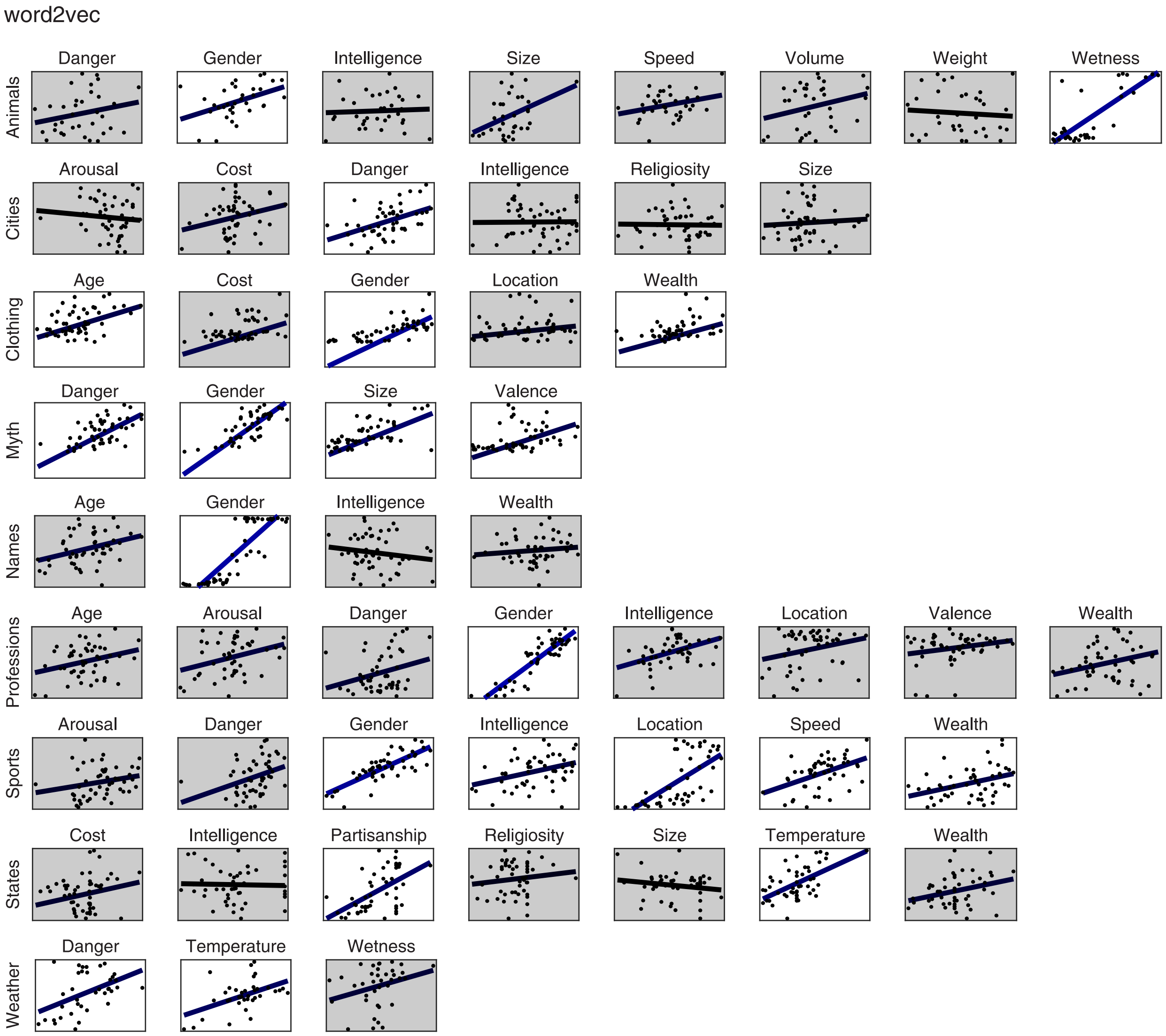
Extended Data Fig. 3 | Correspondence between human judgments and semantic projection using ELMo. Conventions are the same as in Fig. 3 in the manuscript. Descriptive statistics across all tested pairs: (a) Pearson's correlation: med = 0.26 (CI_{95}=0.20-0.36, IQR=0.14-0.43), adjusted med = 0.31 (CI_{95}=0.21-0.41, IQR=0.15-0.45). (b) OC_p: med = 59% (CI_{95}=57-63%, IQR=55-66%), adjusted med = 70% (CI_{95}=65-73%, IQR=63-76%).



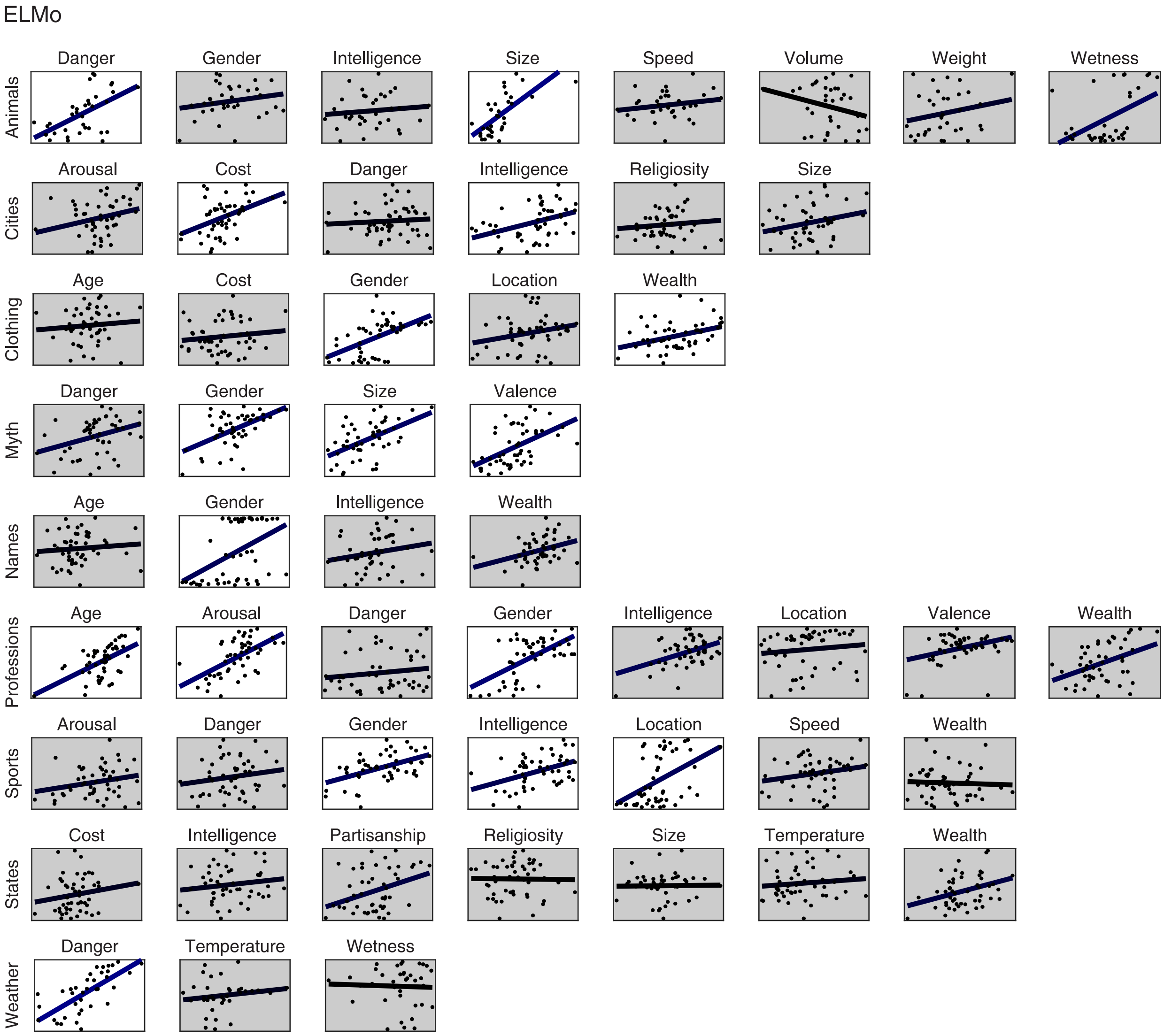
Extended Data Fig. 4 | Correspondence between human judgments and semantic projection using BERT. Conventions are the same as in Fig. 3 in the manuscript. Descriptive statistics across all tested pairs: (a) Pearson's correlation: med = 0.42 (CI_{95} = 0.35-0.47, IQR = 0.20-0.54), adjusted med = 0.44 (CI_{95} = 0.40-0.50, IQR = 0.25-0.57). (b) OC_p: med = 65% (CI_{95} = 62-67%, IQR = 56-72%), adjusted med = 74% (CI_{95} = 72-76%, IQR = 67-80%).



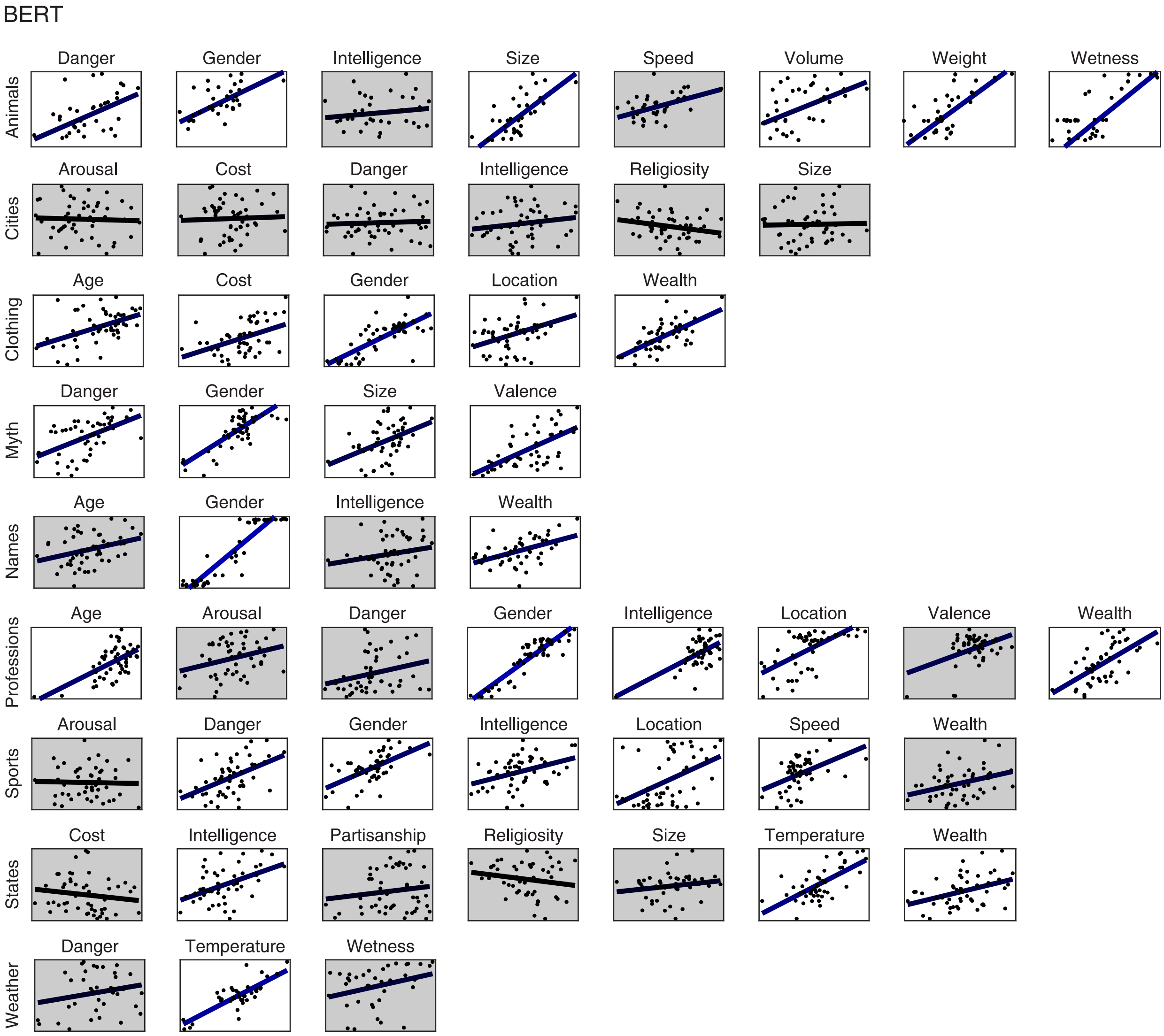
Extended Data Fig. 5 | Detailed results of semantic projection using FastText. Conventions are the same as in Fig. 4 in the manuscript.



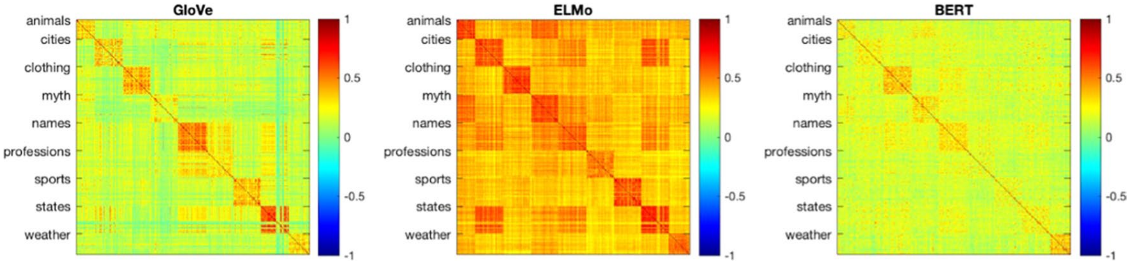
Extended Data Fig. 6 | Detailed results of semantic projection using word2vec. Conventions are the same as in Fig. 4 in the manuscript.



Extended Data Fig. 7 | Detailed results of semantic projection using ELMo. Conventions are the same as in Fig. 4 in the manuscript.



Extended Data Fig. 8 | Detailed results of semantic projection using BERT. Conventions are the same as in Fig. 4 in the manuscript.



Extended Data Fig. 9 | Evaluating how well different word embeddings capture conceptual category structure. Each matrix shows Pearson's correlations between all pairs of word vectors for all items used in our study, grouped by category (indicated on the y-axis), for a different embedding. Color corresponds to correlation strength, with dark blue corresponding to -1 and red corresponding to 1. Qualitatively, all three embeddings capture categorical structure, as is evidenced by the block-diagonal structure of the correlation matrix. Nonetheless, ELMo appears to generate highly similar vectors for words sharing a category (the diagonal blocks are colored in strong red), indicating a poorer ability to distinguish among within-category items, compared to the other two embeddings. In contrast, BERT appears to separate items from across different categories more poorly than the other two embeddings (the color differences between the diagonal blocks and the rest of the matrix are somewhat weak).

nature portfolio

Corresponding author(s): Idan A. Blank

Last updated by author(s): Jan 27, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Behavioral data were collected online, via Amazon Mechanical Turk.
Data analysis	All analyses were carried out in MATLAB, using built-in functions and custom (in-lab) scripts. The only exception is the analysis reported in "Semantic projection is successful even without outlier items", which was carried out in the statistical software R using the lme4 package.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All behavioral data and GloVe vectors as reported in the paper are available on the Open Science Framework (https://osf.io/5r2sz/). The full database of GloVe vectors (including many words not used in this study) is available for download from https://nlp.stanford.edu/projects/glove/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Quantitative behavioral data are fitted with a quantitative models derived from corpora of the English language.
Research sample	For each of 56 experiments, we collected data from n=25 participants (for a total of 1,400 participants) on Amazon Mechanical Turk (MTurk). We limited participation to MTurk users in the United Stated who had previously completed at least 1,000 experiments (“human information tasks” or HITs) with an acceptance rate of 95% or above. Age and sex information was not collected. We measured common-sense knowledge that is, overall, shared across English speakers in (at least) Western, Educated, Industrialized, Rich Democracies. The manuscript provides information about the inter-rater reliability of the judgments we collected, i.e., the extent to which different English speaker "agree" about their common-sense knowledge. Previous studies have established that online data from MTurk replicate in-lab judgments of linguistic materials that far exceed the current ones in complexity / subtlety, which we believe justifies the use of this platform for our purposes.
Sampling strategy	We chose n=25 participants per experiments so that data were approximately normally distributed (following the central limit theorem). We used a convenience sample of MTurk workers.
Data collection	Data were collected online via Amazon's Mechancal Turk platform. Participants completed the experiment (rating different nouns from a given category, e.g., "animals", along some semantic feature, e.g., "size") remotely.
Timing	56 experiments were run between August 2016 and February 2017.
Data exclusions	Exclusion criteria were pre-established and are described in the manuscript. In each experiment, for each participant, we computed the correlation between the ratings they provided and the remaining ratings averaged across the rest of the sample. Participants whose correlation was inferior to the mean correlation in their respective sample by more than 2.5 standard deviations (i.e., participants whose ratings showed weak correlations/agreement with the rest of the group) were removed from further analysis. In the majority of experiments, no participants were excluded, and no more than 2 participants were excluded from any given experiment.
Non-participation	Participants opted-in for the experiment. Each experiment lasted approximately 5 minutes, and no participant dropped out.
Randomization	Participants were not allocated into experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☒	☐	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	See above. Participants self-select (opt-in) for participation. We do not believe the sample thus introduced is biased in terms

Recruitment	of the data collected here, i.e., common-sense knowledge about everyday object categories like animals, items of clothing, professions, etc. Also see data exclusion criteria above.
Ethics oversight	MIT's committee on the use of humans as experimental subjects.

Note that full information on the approval of the study protocol must also be provided in the manuscript.