Local similarity and global variability characterize the semantic space of human languages

Molly Lewis[a], Aoife Cahill[b], Nitin Madnani[b], and James Evans[c,d]

[a]Carnegie Mellon University
[b]Educational Testing Service
[c]University of Chicago
[d]Santa Fe Institute

How does meaning vary across the world's languages? Some claim word meanings are universal, and empirical research has shown cross-linguistic similarity in associations between concrete natural objects.[1] Other research exploring specific domains such as color[2] and kinship[3] has demonstrated variability in the organization of word meanings by language, culture and environment. To what extent word meanings culturally vary or universally agree in different languages remains an unresolved empirical puzzle, but the emergence of powerful word embedding models[4,5] enable us to take a systems-level approach that directly compares languages across semantic domains.[6,7] Here we show that meanings across languages are characterized by similarity within semantic domains and variability across them, using models trained on both (1) large corpora of native language text comprising Wikipedia articles in 35 languages and also (2) English TOEFL essays written by 38,500 speakers from the same distinct native languages, which cluster into semantic domains. The consistency of our findings across these contexts reveals that even successful bilingual communicators think along global semantic associations from their native language[8] while writing English. Consistent with universalist claims, concrete

meanings are less variable across languages than abstract meanings, but consistent with relativist claims, all meanings vary with geographical, environmental and cultural distance. By simultaneously examining local similarity and global difference, our findings harmonize these claims and provide the first description of general principles that govern variability in semantic space across languages. The global structure of a speaker's semantic space influences the comparisons and metaphors cognitively salient to people who natively speak different languages. These findings have dramatic implications for language education, cross-cultural communication, and literal translations, which are impossible not because the objects of reference are uncertain,[9] but because associations, metaphors and stories interlink meanings differently in one language than another.[10]

The answer to questions about the complex relationship between meaning and language strikes to the heart of a fundamental question in the social and communication sciences – are word meanings universal or variable across the world's languages. What precisely is the relationship between the meaning of words "bell", "smell and "hell" in English and their closest translations in Persian or Russian? An emerging body of data collected by anthropologists documents substantial variability across languages regarding their semantic organization within a variety of specific domains, such as color[2] and kinship.[3] On the other hand, there is evidence that in cases of shared, concrete experiences, there is a high degree of semantic similarity across languages.[1]

Understanding the nature and degree of cross-linguistic semantic variability is important because it holds cognitive implications for our ability to learn and switch between languages, just as it pinpoints the pitfalls and potential of intercultural communication around the world. To the extent that language meanings are universal,[11,12] the process of learning a language or translating a document is simply the process of mapping linguistic forms to the same units of meaning. In contrast, to the extent that languages vary in their underlying meaning systems,

the process of learning a language or translating an idea requires not only learning new word forms but also acquiring a rich representation of that system. These two possibilities have very different implications for the tasks of language learning, communicating, translating and even reasoning in different contexts. Most notably, if languages vary in their meanings, it is possible that speakers of different languages think in different ways. This provocative possibility has lead to a long-standing debate centered around case studies of particular semantic domains and languages.[13–15]

Here, we describe and explore semantic variability by examining semantic relationships for distinct languages *across* referential domains, rather than within a single domain. By taking this "system-level" approach,[6,7] we are able to identify domains of relatively greater semantic variability and similarity across languages. More importantly, however, this system-level approach allows us to characterize the macro *structure* of variability in the organization of global meanings. We find that variability in the structure of semantic space across languages is characterized by a high degree of similarity within semantic domains, but significantly more difference in the relationships between those domains. This identifies dramatic differences in the set of cross-domain semantic associations, metaphors and similes that speakers from distinct languages have cognitively available to them.

Examining lexical semantics at the system level presents several methodological challenges. Classic work on cross-linguistic semantics has explored the relationship between words within a single semantic domain, like color.[2,16,17] Researchers have pursued this approach in part because it is unclear how to compare diverse meanings: Red and pink can be compared along dimensions of lightness or saturation, but how does one compare the meaning of red to the meaning of mother? And yet the relative position of diverse meanings condition the space of cognitively available associations. The domain-centric approach is further limited by its requirement that the analyst define relevant semantic domains of inquiry, thereby imposing

idiosyncratic structure and the potential for bias.

Here we address these challenges by taking advantage of a recent advance in machine learning: neural network approaches to word embeddings.[4,5] Word embeddings provide a system-level description of semantics derived from the complex distribution of word collocations in a corpus of text. In the word embedding framework, each word is represented as a high dimensional (e.g., 200) vector, and distance between vectors corresponds to similarity between words, with closer words indicating more similar meanings. Word embeddings have been shown to be highly correlated with human judgments of semantic similarity and to encapsulate and represent culture-specific biases with fidelity.[18–23] We describe computed word embeddings as representing the semantic space of a language and explore semantic distance between pairs of languages in this space by evaluating continuous distances between word pairs in both. We then cluster words based on their loadings on embedding dimensions to compare local (within cluster) versus global (between cluster) variability in semantics between languages.

Using word embeddings, we compare the structure of semantic space for 35 different languages in two stages with two distinct datasets. First, we examine the direct relationship between concrete and abstract words, and between local and global word distances in the context of a large, naturalistic corpus of native language text, an embedding of all Wikipedia entries produced within each language. In Extended Data Figure 1, we show that for the 35 languages we examine, engagement with Wikipedia is comparable in terms of article production and consumption.

Second, we seek to validate these patterns, controlling for differences in topic, lexicon, and syntax, by analyzing TOEFL essays written in English by second language learners from the corresponding languages. This allows us to examine the semantics of different languages while holding constant native language grammar and lexicon. It also allows us to control for broad topic, as all essays are written in response to the same prompts. The striking similarity of

patterns between findings from these two datasets confirm that the semantics of one's native language influences the semantics of one's second language for bilingual speakers,[8] such that language learners from Athens "think Greek" while writing English.

Together these two datasets provide converging evidence about the structure of meaning across human languages. We find substantial variability across languages in the structure of semantic space, but that the relationship between the semantic systems of different languages is predictable. Languages spoken by speakers culturally and geographically more similar manifest comparably similar semantic spaces. Furthermore, we find that the *ways* in which languages differ from each other is principled: Languages tend to vary much more across semantic domains than within them.

Semantic difference between languages

To evaluate the overall semantic differences between languages, we examined the position of TOEFL essays in an embedding space as a function of the native language of the essay writer. We quantified semantic distinctiveness at the language level by taking the difference between mean pairwise cosine distances for essays written by speakers of a particular native language, relative to distances between essays written by different native language speakers. This value was substantially greater than zero for all languages in our sample (M = 0.018, SD = 0.007; $t(34)$ = 16.35, $p < .00001$), suggesting that each language was associated with a distinct semantic space. This difference was also observed in a non-parametric analysis (W = 630, $p <$ .00001). Furthermore, low scoring essays (M = 0.02, SD = 0.006) were more distinct than high scoring essays (M = 0.016, SD = 0.007; $t(34)$ = 3.91, $p < .001$; W = 517, $p < .001$; see ED Fig. 2), suggesting that as learners become more skilled in English, their English semantics diverge from those in their native languages. Nevertheless, high-scoring essays continue to display semantic associations from the native language, and differences were not attributable to

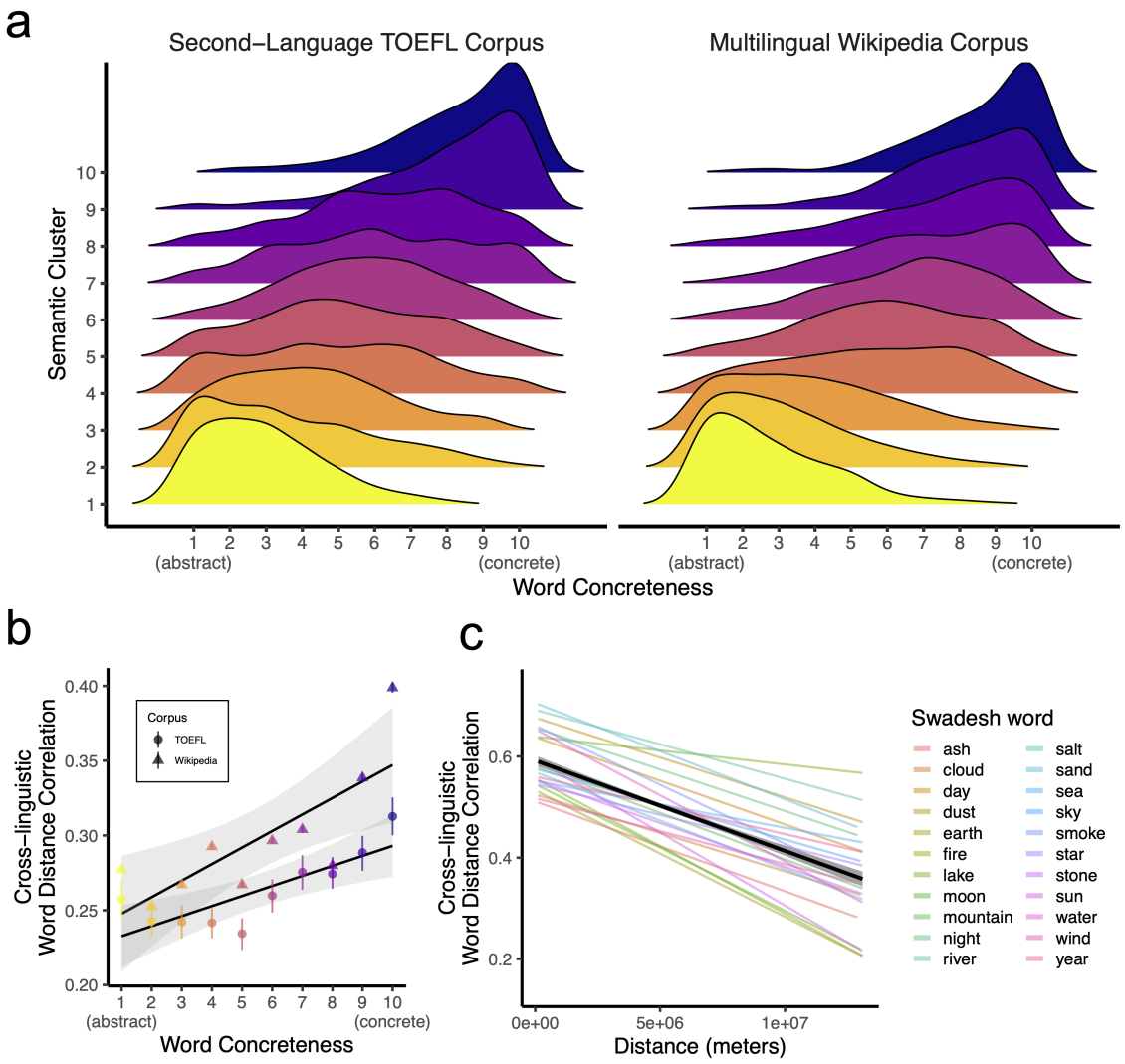

Figure 1: (a) Distribution of words in each semantic cluster across concreteness deciles based on words embeddings obtained from Second-Language TOEFL Corpus and Multilingual Wikipedia Corpus. (b) Mean cross-linguistic word distance correlation (Pearson's r) as a function of the concreteness decile of the words. Larger values indicate more semantic similarity across languages. Point shape indicates corpus. Point ranges correspond to bootstrapped 95% confidence intervals; Range on model fit corresponds to the standard error. (c) Linear model fits for cross-linguistic word distance correlation (Pearson's r) as a function of the geodesic distance between two languages (meters). Each data point corresponds to a unique language-pair-word combination. The colored lines correspond to the model fit for each word, and the black line shows the overall model fit and the corresponding standard error.

English grammatical or syntactical errors (see ED Fig. 3a).

Concrete concepts translate better than abstract ones

Having validated second-language text as a method for analyzing cross-linguistic semantic variability, we next examined cross-linguistic similarity in the structure of semantic space. Universalists argue that all languages share roughly the same semantic structure, while relativists argue for substantial variability.

We hypothesized that the *amount* of variability for a particular semantic domain across languages would vary, but in a principled manner. We reasoned that semantic domains referring to meanings that were more *perceptually* available and concrete such as "food" and "body" would be more similar across languages, relative to domains more conceptual and abstract like "injustice" and "democracy". This hypothesis is motivated by the idea that, while there is substantial variability in the cultures and environments in which languages are spoken, all speakers share roughly the same perceptual systems and would therefore be more likely to experience similar concrete objects in similar ways.

To test this hypothesis, we estimated the concreteness of each word based on human judgments, and partitioned them into 10 contiguous sets separated by rising concreteness thresholds.[24] These sets strongly overlapped with semantic clusters in both the TOEFL ($\chi^2(81) = 1538.1$; $p < .00001$) and Wikipedia word samples ($\chi^2(81) = 5144.1$; $p < .00001$; Fig. 1a). In line with universalist claims, languages exhibit higher similarity in more perceptually concrete domains, and less in those more conceptually abstract (TOEFL: $r = .78$; $p = .008$; Wikipedia: $r = .82$; $p = .004$; Fig. 1b; see ED Figs. 4-5 for supporting analyses).

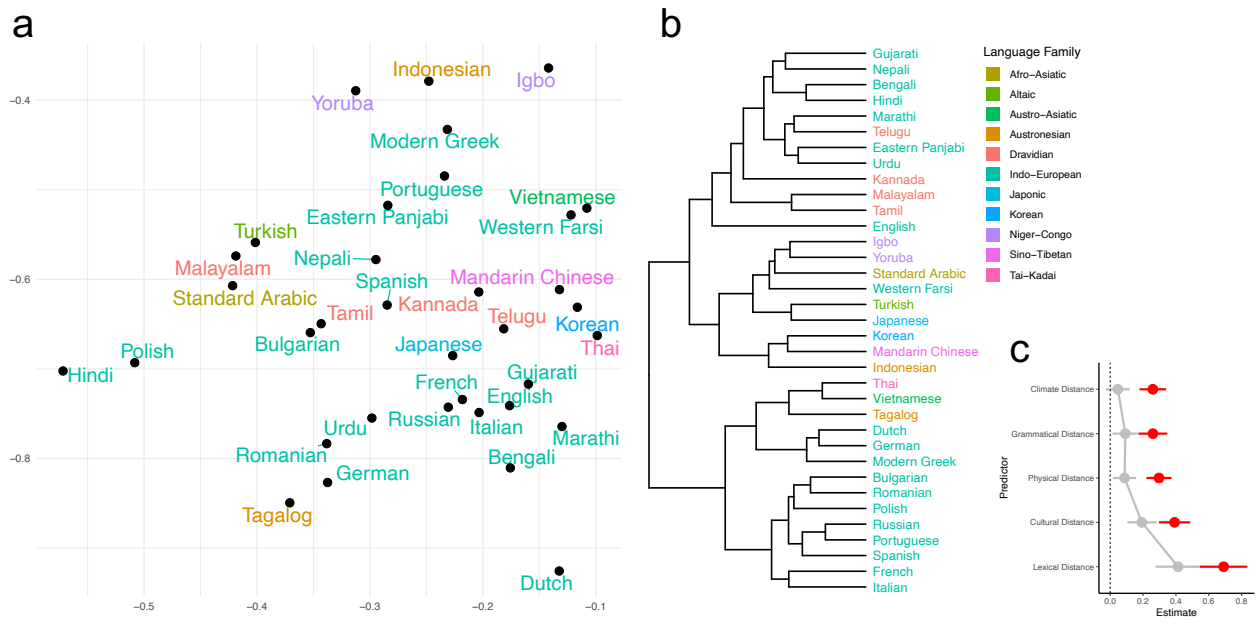

Figure 2: (a) Two-dimensional projection of language centroids calculated from document embeddings of the Second Language TOEFL Corpus. Color corresponds to language family. (b) Hierarchical clustering of languages based on pairwise language distances of language centroids.; (c) Predictors of semantic distance. Ranges are 95% confidence intervals. Red points indicate estimates from single-predictor model; grey points indicate estimates from additive linear model with all five predictors included.

.

Environment and culture predict semantic deviations

Even within highly concrete domains, however, we observed appreciable variability in the structure of semantic space across languages. We estimated pairwise-distances between the 22 primitive words examined by Youn et al. (2016; e.g. “water”, “sun”, “dirt”), and still found moderate variability in pairwise-distances across languages.

This variability, however, was highly predictable by physical and environmental distances. Languages in closer physical distance *much* more similar semantic representations for almost all of these highly concrete words (QAP $p < .01$: physical: 20/22; Fig. 1c; ED Fig 6) and environmental distance explains variations for some of the items (environmental: 6/22). This

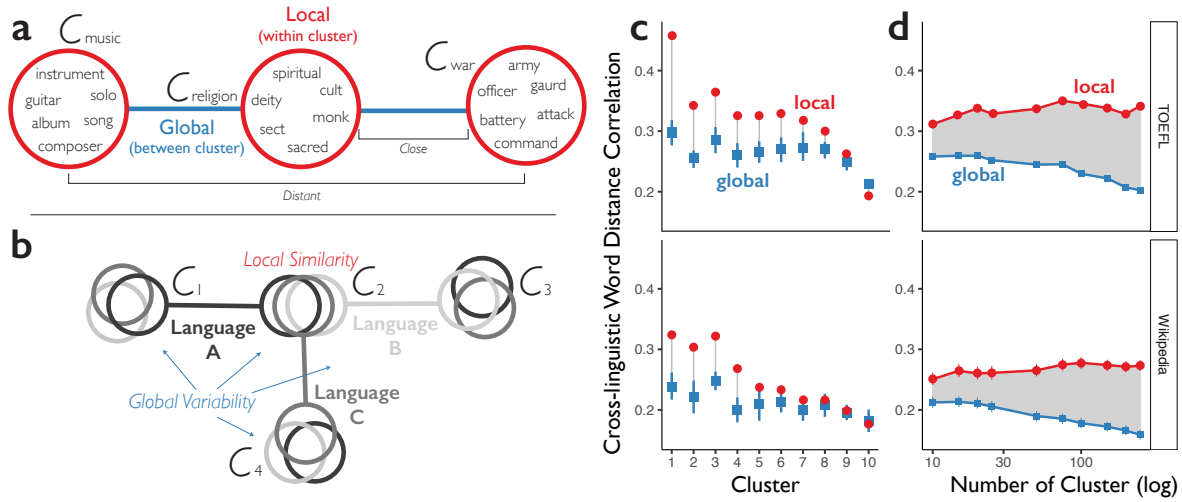

Figure 3: (a) Schematic representation of local and global distances in word embedding models. The figure shows three "clusters" of meanings and the relationships between them. Local relationships are within-cluster distances; global relationships are between-cluster distances. Some clusters are closer globally (e.g., C_{music} and $C_{religion}$) than others (e.g., C_{music} and C_{war}). (b) Schema of the structure of cross-linguistic variability in local and global semantic structure. Each shading corresponds to a language. Languages tend to have overlapping clusters–to share local similarity–but vary in their global relations (e.g. C_1 and C_2 are globally close in Language A, but not in Language B). (c) Cross-linguistic word distance correlations for word groups in 10 semantic clusters based on words embeddings obtained from Second-Language TOEFL Corpus (top) and Multilingual Wikipedia Corpus (bottom). Red points indicate mean local correlation for each cluster, and blue squares indicate global correlation for each cluster. (d) Cross-linguistic word distance correlations for local versus global semantic comparison as a function of the number of semantic clusters. Point ranges correspond to 95% confidence intervals.

suggests that even for concrete meanings, there is variability in the structure of semantics across languages and this variability is predictable by a combination of differences in the perceptual experience of those language speakers and the potential for direct or indirect cultural contact.

When we seek to explain the similarity of semantic structure for the entire lexicon across languages as a function of similarity between language contexts, many dimensions of similarity exercise independent significant influences. These include geographical and environmental similarity, phonological and grammatical similarity, and cultural similarity comprising factors

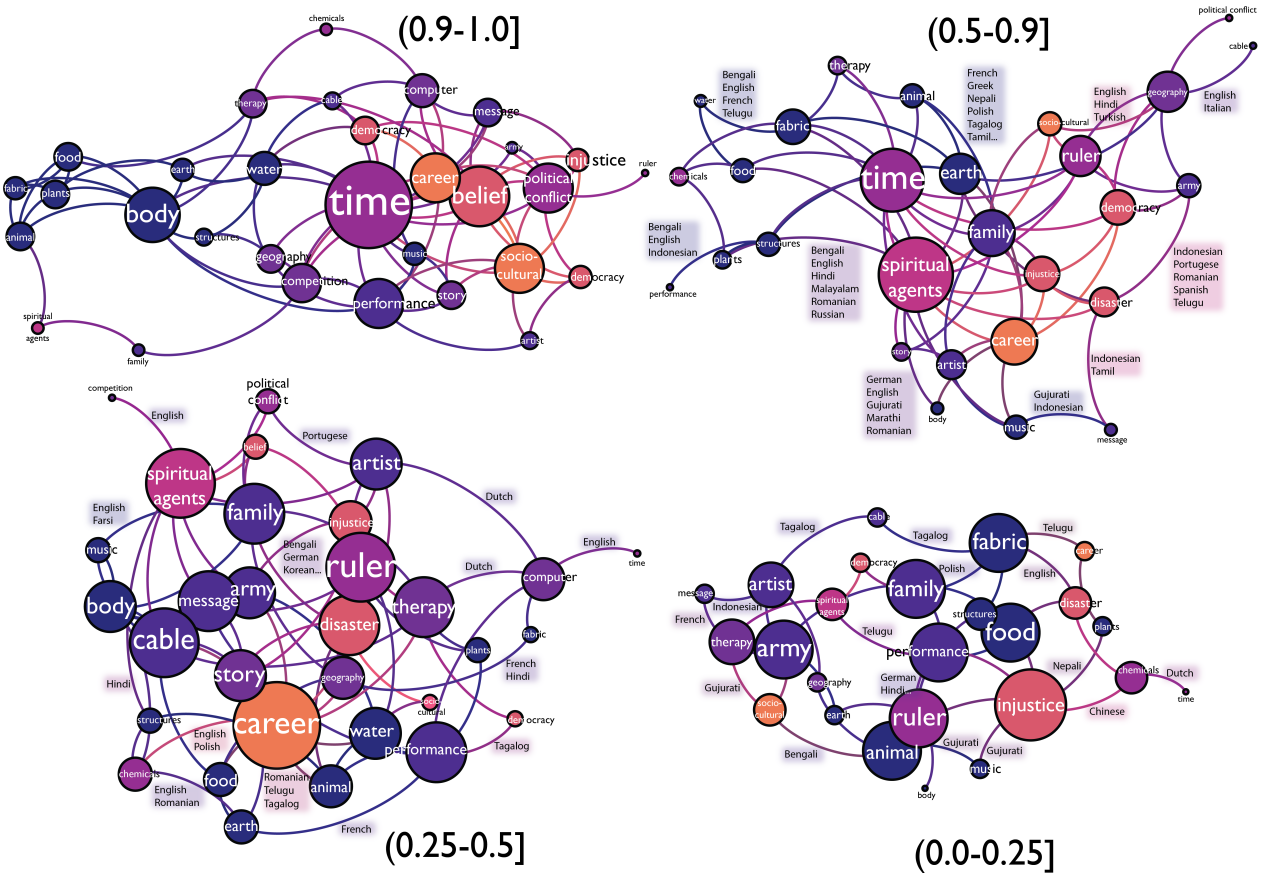

Figure 4: Semantic cluster graphs where links between clusters represent whether or not those clusters are within the top quartile of global semantic closeness for the proportion of languages that share both clusters specified above each graph. Links between clusters in the top left are shared by almost all (0.9-1.0] languages in which the clusters manifest. Graphs in other frames represent global associations shared among fewer languages (0.5-0.9], (0.25-0.5], and (0.0-0.25]. These differences illustrate that despite stable, shared local semantic clusters across languages, global associations between those clusters vary dramatically. All graphs are rendered in 2D using force-directed algorithms that draw together the most connected clusters. See more details about data, cluster labels, and 2D layout in the Methods and Extended Data below.

ranging from likeness in the structure of kinship, religion, politics and social class (Fig. 2; ED Figs. 7-8). Contradicting claims of universal semantics, these analyses suggest that language semantics vary across languages, but remain predictable by cultural difference and environmental distance.

Local similarity and global variability in semantic space

These analyses motivate us to examine how languages vary in their structure at the "system" level across semantic domains. Across the semantic system, word meanings may differ in terms of their *local* semantic relations within a semantic domain – e.g., the relative similarity of meanings associated with "earth" and "sun" (Fig. 3). Alternatively, meanings may differ in terms of their *global* semantic relationships across semantic domains – e.g. the relative similarity of the "astronomy" cluster of meanings ("earth," "sun," etc.) to the "religion" cluster ("spirit," "hell," etc.). Finally, meanings may differ evenly across the system. We first found that word pairs from the same concreteness decile ("local" relations) were more similar to each other across languages, relative to word pairs across concreteness deciles ("global" relations; TOEFL: $M = 0.025$, $SD = 0.005$; $t(594) = 128.64$; $p < .0001$; d = 1.28 [1.15, 1.4]; W = 177310, $p < .0001$; Wikipedia: $M = 0.035$, $SD = 0.015$; $t(594) = 56.88$; $p < .0001$; d = 0.27 [0.16, 0.39]; W = 177272, $p < .0001$). To more directly measure local-global semantic relations, which are not exclusively organized by concreteness, we examined the relative contribution of these two types of variability for word groups defined by semantic clusters (see Methods for detail). Critically, we found that local semantic clusters were much more correlated across languages, relative to global semantic structure, and demonstrated stronger differences than between concreteness deciles in both the TOEFL ($M = 0.058$, $SD = 0.008$; $t(594) = 185.97$; $p < .0001$; d = 2.84 [2.68, 3]; W = 177310, $p < .0001$) and Wikipedia datasets ($M = 0.038$, $SD = 0.024$; $t(594) = 38.27$; $p < .0001$; d = 0.31 [0.19, 0.42]; W = 174402, $p < .0001$; Figs. 4a-5; ED Figs. 9-10). This effect was not related to the grammatical similarity of the languages (ED Fig. 3b), and grew substantially larger as the number of semantic clusters increased (Fig. 4b).

Implications for language learning and cross-cultural communication

These findings highlight limitations to binary characterization of the world's languages by semantic universality or semantic relativity. Consistent with semantic universalists who demonstrate semantic similarity between concrete, "natural" concepts, we show that the configuration of concrete concepts in semantic space is much more likely to be conserved across languages than abstract concepts such as democracy, tragedy, or spiritual agents like angels and demons. Consistent with semantic relativists, we show that the distance between concepts in semantic space is a function of lexical and grammatical distance between the languages, and the cultural and physical distance between the peoples and place where they are spoken. Neither semantic universalism nor relativism, however, characterizes the most significant, overarching pattern of similarity and difference across languages. Languages tend to be locally similar and globally varied: they are similar in how they cluster words with meanings proximate to one another, but divergent in how those clusters relate across the entire semantic space. Words associated with foods, body parts, spiritual agents, and tragedies individually tended to cluster together, but relations between them varied widely by language.

The global structure of a speaker's semantic space necessarily influences the comparisons and metaphors cognitively salient to people who natively speak different languages: People who "think" in Greek cognitively follow and produce distinct semantic associations in text from others who "think" in Arabic, Farsi, Igbo or Chinese. This variability has powerful practical implications for language learning – it suggests that learning words in a new language is not just a process of learning word forms and their mappings to referents, but also the higher-order association between their meanings. Currently, second-language training begins with explicitly translated word associations, often clustered locally within domains of experience (e.g., words

for objects found in a house), interleaved with grammatical patterns required to use those words correctly in sentences. Later, language learners are introduced to global associations between clusters of meaning implicitly through native language literature and traditional stories. When people from a language write about family, do they also tend to link it with concepts of immutable stones and mountains, health, tragedy, or (in)justice? Our findings demonstrate that global associations could also be explicitly represented and taught. Moreover, they suggest the importance of recognizing that to speak with native fluency within a language, one needs to "think" in that language, producing global associations familiar to that language culture.

Variability in global semantics has dramatic implications for cross-cultural communication and collaboration. It suggests that faithful, word-by-word translations are not possible, not because the objects of reference are uncertain,[9] but because associations, metaphors and stories interlink different domains of meanings in one language culture than another.[10] This means that communication between two people of different language backgrounds will necessarily lead to some loss and distortion of intended meaning. Further, it points to the intriguing possibility that communication will be more faithful among speakers of semantically more aligned languages: A native speaker of Turkish can more effectively communicate in a second language with a native speaker of a semantically similar language, like Japanese, compared to a dissimilar one, like Dutch. Moreover, this suggests the collective cognitive diversity that might emerge from collaboration among those from native languages with very different semantic structures. Understanding the ways in which languages differ in their semantics has the potential to facilitate better cross-cultural communication, and provide a new justification of the importance of cultural difference. Our work is the first to describe the general principles that govern this variability and to characterize these differences across semantic space.

Methods

Corpora and Models

The Second Language TOEFL corpus contained 38,500 short essays written in English by second-language learners of English. Each essay was written in response to one of 28 different essays prompts. The essays were written by equal number of participants from 35 different languages. Each essay was associated with a 1-5 score, implying an essay that ranged from poor to excellent.

To evaluate whether each language was associated with a distinct semantic space, we trained a single `doc2vec` model[25] on this corpus with the output vector of 200 dimensions and a window size of 6. We used the gensim implementation of the `doc2vec` model.[26] For each target language, we then sampled 100 essays from each language and estimated the mean cosine distance within all essays in the target language and the mean cosine distance between essays in the target language and essays in other languages. We repeated this procedure 100 times for each language and estimated the mean within and between cosine distance in each language. To quantify the semantic distinctiveness of each language, we calculated the difference (and, alternatively, the ratio, see ED Fig. 2) of within to between essay distances. This value should exceed zero for differences (or one for ratios) if each language is associated with a distinct semantic space. We report both parametric (t-test) and non-parametric (Wilcoxon signed-rank test) analyses of the overall distinctiveness of essays by language, and within-language comparisons of distinctiveness for low versus high scoring essays. All tests are two-tailed.

All remaining analyses were performed on models trained on each language of the 35 languages separately. Multilingual Wikipedia models were trained on corpora of Wikipedia articles in each of the target native languages using the `word2vec` skip-gram algorithm with default parameters.[4] The Second Language TOEFL models were trained on 35 corpora separated by

the native language of the essay writer using the same training parameters as above (Figs. 1b, 4-5).

Word level analyses

The conceptual concreteness of a word was estimated using previously-collected human judgments.[24] Participants were presented with a single English word, and asked to rate the conceptual concreteness of its meaning on a 5-pt Likert scale, ranging from abstract to concrete. The notions of concreteness and abstractness were defined for participants as follows: "Some words refer to things or actions in reality, which you can experience directly through one of the five senses. We call these words concrete words. Other words refer to meanings that cannot be experienced directly but which we know because the meanings can be defined by other words. These are abstract words." Judgments were collected for a sample of 39,954 words.

For analyses using the Multilingual Wikipedia Corpus, we translated all words in the concreteness dataset into each of the target 35 languages using the Google Translate API. We selected the set of words that had translations for at least 30 of the languages, and then sampled 1,000 words from each of decile of concreteness (based on the human judgments described above). Of our target sample of words, 45% of the translations existed in the embedding models across all languages. For the Second Language TOEFL corpus, we selected all words that were present in the models of 5 or more languages (N = 3,530 words). The words in this sample were roughly uniformly distributed across deciles. Each word in our sample was rated for concreteness by at least 21 participants (TOEFL: $M = 54.9$, $SD = 398$; Wikipedia: $M = 37.5$, $SD = 236$), and there was high agreement across participants in their rating of conceptual concreteness (TOEFL: Mean SD across words $= 1.19$; Wikipedia: $SD = 1.15$).

We compared word sets defined by different levels of concreteness to word sets defined by semantics. For both the Wikipedia and TOEFL models, we used k-means clustering[27] to cluster

the words into 10 clusters each based on their semantics. Clusters were determined based on the model trained on English Wikipedia. We report the χ^2 statistic of word counts in a N-cluster by concreteness decile matrix (10 x 10).

We next evaluated the semantic similarity of words across languages as a function of word concreteness. We calculated the pairwise distance (cosine) between all words within each concreteness decile. We then calculated the correlation for these word distances for each language pairing (N = 595). Finally, we averaged across language pairs to obtain an estimate of the mean cross-linguistic correlation in word distances across languages for each decile. Correlation values are Pearson's r.

To characterize cross-linguistic differences in local versus global similarity, we compared the pairwise cosine distances between words in different concreteness deciles ("global"), to those in the same concreteness decile ("local", described above). Using the same set of words as above, we measured the pairwise distance between words in different concreteness deciles, and then calculated the correlation for each language pairing and decile pairing (1-2, 1-3, 1-4, etc). We averaged across decile pairs of the same local-global type, and then compared local and global distances for each language pair. We report statistics for both parametric (paired t-test) and non-parametric (paired Wilcoxon signed-rank test) analyses. A parallel analysis was also conducted using word sets defined by the semantic clusters described above, varying the number of clusters considered (10 - 250). Means and standard deviations presented for these analyses correspond to the difference in correlation between local and global distances. Effect size measures are Cohen's d and corresponding 95% confidence interval.

Semantic similarity in Swadesh words

We used the Google Translate API to translate the 22 words analyzed by Youn et al. (2016)[28] (a subset of the Swadesh list) into each of our target 35 languages. We included the variants an-

alyzed by Youn et al. (e.g., "day"/"daytime", "ash"/"ashes"), averaging across words referring to the same concept. We obtained translations for 96% of the words across languages using this method. We then used these translations to obtain embedding coordinates for each concept in each language from the Wikipedia-trained embedding model.[4] In cases where translations were available for multiple word forms (e.g., "day" and "daytime") or the translations were composed of multiple forms, we averaged across vectors. We calculated the pairwise distance (cosine) between each unique word pair (231 pairs) in each language. Then, for each word, we estimated the correlation (Pearson's r) between these distances for each language pair (595 language pairs). We estimated the physical distance between languages by obtaining the geographical coordinates of each language from Glottolog 2.7[29] and calculating the geodesic distance (distance on an ellipsoid) between each language pair. Finally, we correlated the language-pairwise distance correlation coefficient with the language-pairwise physical distance metric and estimated p-values using the Quadratic Assignment Procedure (QAP)..[30] The QAP procedure estimates p-values in a way that accounts for the non-independence of observations (see methods below).

Climate similarity was based on climate data obtained from WorldClim[31] on the basis of geographical coordinates from Glottolog 2.7.[29] For each language pair, we measured the Euclidean distance between estimates of mean and variance in temperature and precipitation. Measures of linguistic distance were obtained from.[32] Typological distance between languages is based on similarity of 130 typological features for each language coded from the WALS database.[33] Phonological similarity is based on the Levenshtein edit distance between a standard set of 40 words in each language[34] (ASJP16). Finally, cultural similarity is based on data from D-place, an ethnographic atlas of cultural traits.[35,36] The cultural distance measure presented in Fig. 2c is an aggregate measure of cultural traits from 10 domains ("agriculture and vegetation," "actions and technology, "emotions and values," "kinship," "law," "possession," "religion and belief," "social and political relations," "the house," and "the physical world.").

See ED Figure 7b for by-domain analyses.

Cosine distance as similarity metric

Similarity and distance between words in an embedding space is typically assessed using "cosine similarity," the cosine of the angle between two word vectors ("cosine distance" is one minus the cosine between vectors). This is preferred to the Euclidean (straight-line) distance due to properties of high-dimensional spaces that violate intuitions formed in two or three dimensions.[22] For example, as the dimensionality of a hypersphere grows, its volume shrinks relative to its surface area as more of that volume resides near the surface. The surface area of a unit circle surpasses its volume in three dimensions, but as the hyperspheres dimension approaches infinity, its volume approaches zero.

A geometric interpretation may be preferable to a probability one like the Kulback-Leibler Divergence or Wasserstein Distance because the distance between two probability distributions assumes independence and equal weight between each dimension, which is not the case for neural models like `word2vec` that approximate factorization of a (very) large matrix, with a monotonically decreasing influence of each dimension in describing the overall variation of the matrix.[37]

QAP non-independence

Multiple regression quadratic assignment procedures (MR-QAP) tests are permutation tests for multiple linear regression model coefficients for data organized in square matrices of relatedness among n objects.[38] This data structure has been most common in studies of social networks, where variables indicate a relation between n actors, but are equally applicable here, where we explore a distance relationship between n languages. In both network, and distance cases, the rows and columns are explicitly not independent of one another, and so assumptions

of identically and independently distributed data, required for linear regression are misplaced. MR-QAP permutation tests allow us to demonstrate that the autocorrelation among language pairs does not influence the regressed association that we find–that the distances between clusters is significantly more variable than the distances within clusters (e.g., the alpha estimated in the regression of global distances on local distances is significantly greater than 0.) It is common to have coefficients that look highly significant under a classical null hypothesis test and that remain insignificant under MR-QAP because the QAP null hypothesis accounts for autocorrelation.

References

1. Youn, H. *et al.* On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences* **113**, 1766–1771 (2016).

2. Berlin, B. & Kay, P. *Basic color terms: Their universality and evolution* (Univ of California Press, 1991).

3. Murdock, G. P. Kin term patterns and their distribution. *Ethnology* **9**, 165 (1970).

4. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information (2016).

5. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space (2013).

6. Saussure, F. Course in General Linguistics (Peter Owen, London, 1916, 1960).

7. Lévi-Strauss, C. *Structural Anthropology* (Basic Books, 2008).

8. Ameel, E., Malt, B. C., Storms, G. & Van Assche, F. Semantic convergence in the bilingual lexicon. *Journal of Memory and Language* **60**, 270–290 (2009).

9. Quine, W. V. O. *Word and object* (MIT press, 2013).

10. Lakoff, G. Women, fire and dangerous things: What categories reveal about the mind. *Chicago: University of Chicago* (1987).

11. Wierzbicka, A. *Semantics: Primes and universals: Primes and universals* (Oxford University Press, UK, 1996).

12. Fodor, J. A. *The language of thought*, vol. 5 (Harvard University Press, 1975).

13. Winawer, J. *et al.* Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences* **104**, 7780–7785 (2007).

14. Gennari, S. P., Sloman, S. A., Malt, B. C. & Fitch, W. T. Motion events in language and cognition. *Cognition* **83**, 49–79 (2002).

15. Brown, P. & Levinson, S. C. "Uphill" and "downhill" in Tzeltal. *Journal of Linguistic Anthropology* **3**, 46–74 (1993).

16. Regier, T., Kay, P. & Khetarpal, N. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences* **104**, 1436–1441 (2007).

17. Baddeley, R. & Attewell, D. The relationship between language and the environment information theory shows why we have only three lightness terms. *Psychological Science* **20**, 1100–1107 (2009).

18. Hill, F., Reichart, R. & Korhonen, A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* **41**, 665–695 (2015).

19. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).

20. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* **115**, E3635–E3644 (2018).

21. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 4349–4357 (2016).

22. Kozlowski, A. C., Taddy, M. & Evans, J. A. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review* **84**, 905–949 (2019).

23. Lewis, M. & Lupyan, G. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behavior* (in press). URL `https://psyarxiv.com/7qd3g`.

24. Brysbaert, M., Warriner, A. B. & Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* **46**, 904–911 (2014).

25. Le, Q. & Mikolov, T. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, 1188–1196 (2014).

26. Rehurek, R. & Sojka, P. Gensim–Python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* **3** (2011).

27. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **28**, 100–108 (1979).

28. Swadesh, M. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* **96**, 452–463 (1952).

29. Hammarstrm, H. & Nordhoff, S. Langdoc: Bibliographic infrastructure for linguistic typology. *Oslo Studies in Language* **3**, 31–43 (2011). URL `https://www.journals.uio.no/index.php/osla/article/view/75/199`.

30. Butts, C. T. *sna: Tools for Social Network Analysis* (2016). URL `https://CRAN.R-project.org/package=sna`. R package version 2.4.

31. Fick, S. E. & Hijmans, R. J. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* **37**, 4302–4315 (2017).

32. Dediu, D. Making genealogical language classifications available for phylogenetic analysis: Newick trees, unified identifiers, and branch length. *Language Dynamics and Change* **8**, 1 – 21 (2018).

33. Dryer, M. S. & Haspelmath, M. (eds.) *WALS Online* (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013). URL `http://wals.info`. `http://wals.info`.

34. Wichmann, S. *et al.* The ASJP database (version 16) (2013). URL `http://asjp.clld.org/`. `http://asjp.clld.org/`.

35. Kirby, K. R. *et al.* D-place: A global database of cultural, linguistic and environmental diversity. *PLoS One* **11**, e0158391 (2016).

36. Thompson, B., Roberts, S. & Lupyan, G. Quantifying semantic similarity across languages. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (2018).

37. Levy, O. & Goldberg, Y. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, 2177–2185 (2014).

38. Dekker, D., Krackhardt, D. & Snijders, T. A. Sensitivity of MRQAP tests to collinearity and autocorrelation conditions. *Psychometrika* **72**, 563–581 (2007).

Author Information Correspondence should be addressed ML (mollylewis@cmu.edu) or JE (jevans@uchicago.com).

Author Contributions JE and ML designed the research; ML analyzed all data; AC and NM ran the TOEFL models. ML and JE wrote the manuscript.

Acknowledgements We would like to thank the National Science Foundation #1520074 to the University of Chicago for partial funding for this project.