PNAS
www.pnas.org

Supplementary Information for

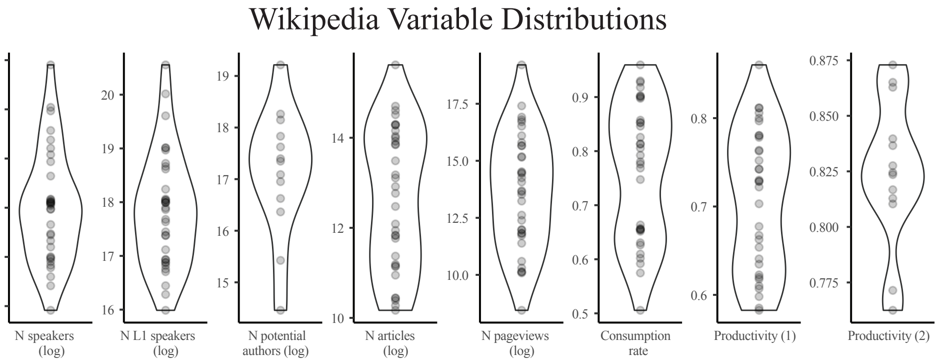
Local similarity and global variability characterize the semantic space of human languages

Lewis, M., Cahill, A., Madnani, N., Evans, J.

Corresponding authors: Molly Lewis (mollylewis@cmu.edu) or James Evans (jevans@uchicago.edu)

This PDF file includes:

Figs. S1 to S21
Table S1
SI References



Wikipedia Variable Correlations

	N speakers	N L1 speakers	N potential authors	N articles	N pageviews	Consumption rate	Productivity (1)
N L1 speakers	0.97 (< .01)						
N potential authors	0.77 (< .01)	0.82 (< .01)					
N articles	0.47 (0.01)	0.47 (0.01)	0.83 (< .01)				
N pageviews	0.50 (< .01)	0.53 (< .01)	0.86 (< .01)	0.95 (< .01)			
Consumption rate	0.20 (0.27)	0.23 (0.18)	0.27 (0.36)	0.91 (< .01)	0.95 (< .01)		
Productivity (1)	0.02 (0.90)	−0.01 (0.94)	−0.13 (0.67)	0.88 (< .01)	0.80 (< .01)	0.92 (< .01)	
Productivity (2)	−0.79 (< .01)	−0.79 (< .01)	−0.64 (0.02)	−0.12 (0.71)	−0.38 (0.20)	0.35 (0.24)	0.74 (< .01)

Fig. S1. *Top*: Distributions for eight descriptive statistics associated with multi-lingual Wikipedia corpora. N speakers indicates the log number of first and second language speakers (N languages missing = 1) (1); N L1 speakers indicates the log number of first language speakers only (N missing = 1) (2). N potential authors estimates the log number speakers of the target language who have access to the internet, and have sufficient command of the target language to author an article (N missing = 22) (3). N articles indicates the log number of articles by language (31 March 2015; N missing = 1) (1). N pageviews indicates the log number of pageviews by language for a 24 hour period (1 August 2016) (4). Consumption rate is the log number of articles / log number of first language speakers (N missing = 1). Productivity (1) is the log number of articles / log number of first language speakers (N missing = 2). Productivity (2) is the log number of articles / log number of potential authors (N missing = 22). Each point corresponds to a language. *Bottom*: Pairwise correlations between Wikipedia measures (Pearson's r). Parenthetical numbers indicate p-values. Critically, languages with more speakers do not have greater rates of engagement, as measured by number of articles written and viewed relative to the number of speakers. In fact, by one measure of article production (Productivity (2)), languages with more speakers write relatively fewer articles per speaker capita.

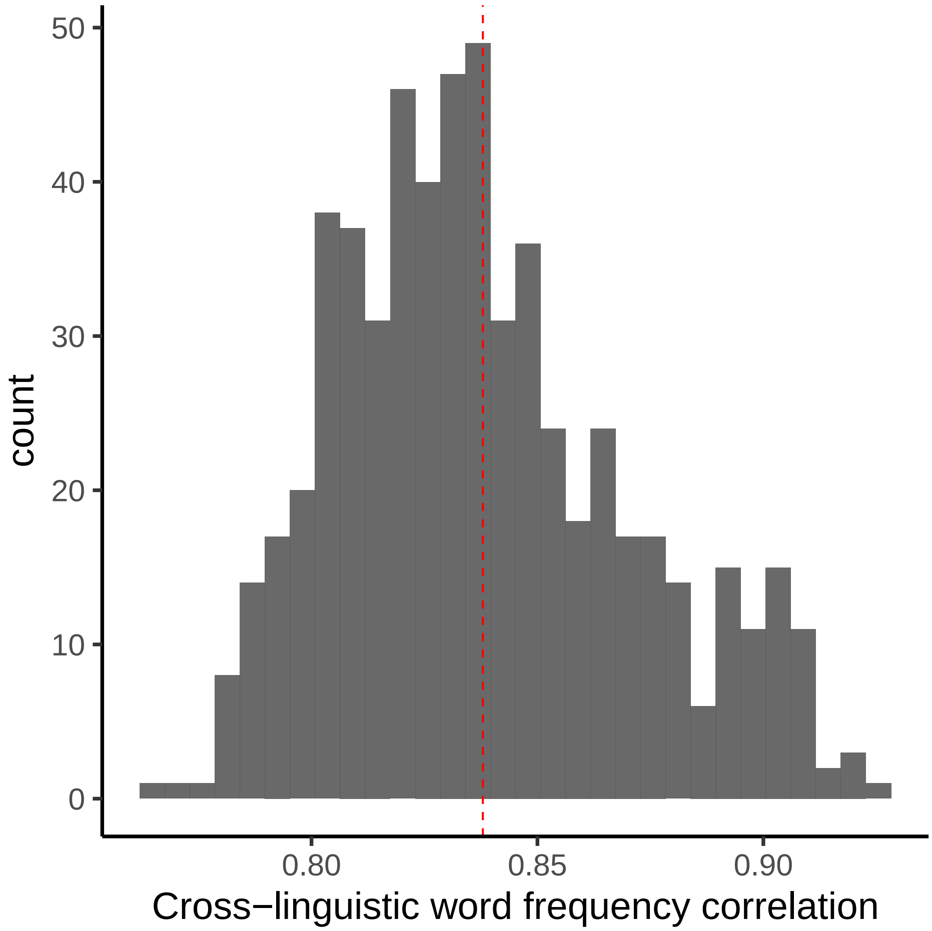


Fig. S2. Are findings driven by differences in languages or what speakers of those languages choose to talk about? When we take Spearman's ρ (rank order correlation) of word frequencies across the ETS essays written by speakers of 35 languages (and limited to words occurring in 10 or more essays across the 38,500 essay sample), we get .84 (for 595 pairs), strongly suggesting that these essays discuss the same topics and broad content, but are hyperlinked with different associations and metaphor as demonstrated by our core findings, illustrated in Fig. 3.

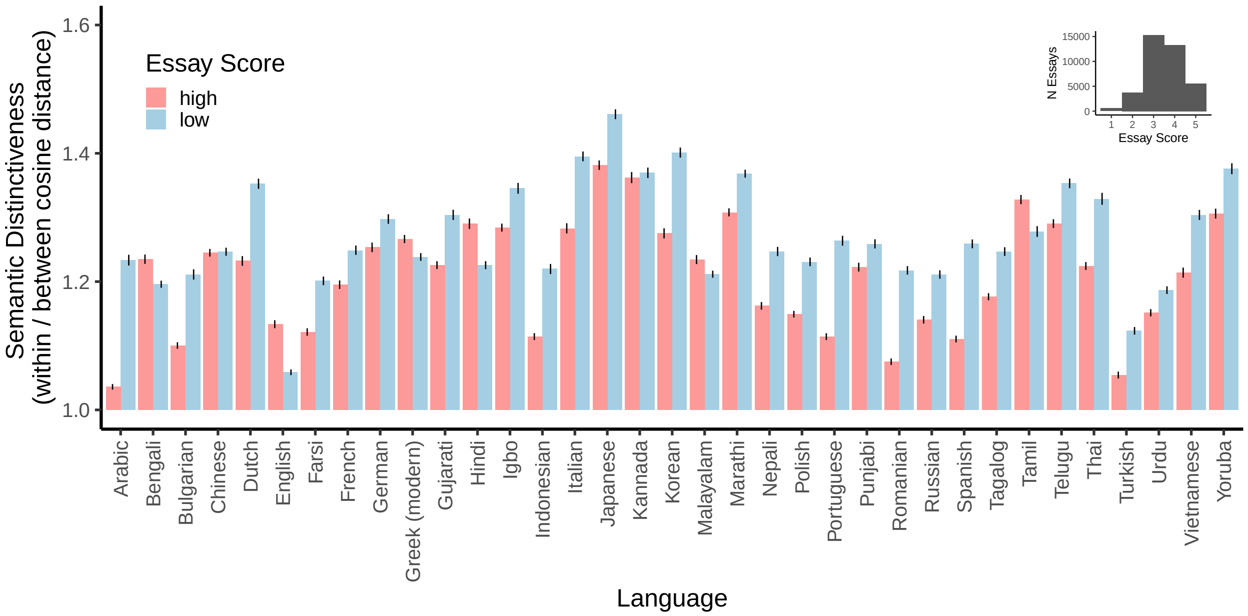
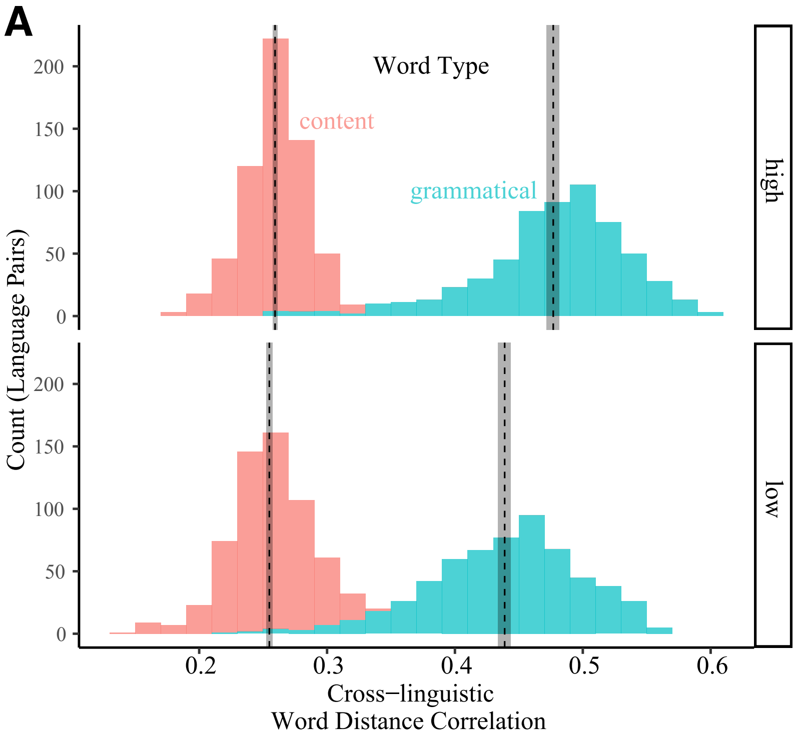


Fig. S3. Essays written in English by speakers of the same native language are more semantically similar than essays written by speakers of different native languages, corroborating informal observations by second language educators and bilingual researchers in individual pairs of languages second language speakers appear to "think" in their first language (5, 6). We evaluated language semantic distinctiveness by analyzing the position of essays in semantic space for a model trained on all essays in all languages. For each language, we averaged the cosine distance between essay pairs from the same language ("within"), and averaged pairs between different languages ("between"). We then quantified language-level semantic distinctiveness in two ways: (1) the difference between the two measure (within - between) as in (7–10), and (2) the ratio between them (within/between) as in (11–13). Both types of measures have been shown to capture meaningful semantic relationships in embedding models. In the Main Text, we report the results for the difference measure; here we replicate our results with the ratio measure. The ratio is greater than one when languages are located in a distinct semantic space, relative to other languages. This value was substantially greater than one for all languages in our sample (M = 1.26; SD = .09; t(34) = 17.09; $p <$ .00001). We also conducted this analysis for two types of essays separately: low scoring essays (score $<$ 4 on 5 pt. scale) and high scoring essays (score $\geq$ 4). Low scoring essays (M = 1.27; SD = .08) were more distinct than high scoring essays (M = 1.21; SD = .09; t(34) = 5.73; $p <$ .00001). Main panel shows mean distinctiveness value across samples with the ratio measure. Red bars correspond to high scoring essays; blue bars correspond to low scoring essays. Ranges indicate bootstrapped 95% confidence intervals. Upper right panel shows the distribution of essay scores (1-5) for the 38,500 essays in the TOEFL corpus (M = 3.51; SD = .91).



B

term	estimate	SE	t-statistic	p-value
(Intercept)	0.04	0.004	10.542	$<.0001$
Grammatical Distance	$<.0001$	$<.0001$	-0.627	0.531

term	estimate	SE	t-statistic	p-value
(Intercept)	0.058	0.001	47.725	$<.0001$
Grammatical Distance	$<.0001$	$<.0001$	0.583	0.56

Fig. S4. A. Grammatical words (e.g., “and", “the") are more similar across languages than content words (e.g., “love"). In Jakobson’s functional typology (14), grammatical words are *metalingual*–they govern the structure of the code of language itself and so play a related function across languages. As such, grammatical words are qualitatively different from abstract content words such as “career" or “government" that vary based on distinctive complex cultural environments. We divided the words from the TOEFL essays analyzed in the Main Text (N = 3,530) into two sets – “content" and “grammatical" – based on part of speech information from (15). Nouns, verbs, adverbs, adjectives, or interjections were categorized as content (N = 3,339); articles, determiners, conjunctions, pronouns, or prepositions (N = 145) were categorized as grammatical. We then calculated the cosine distance between each word pair within a word type (content or grammatical) for models trained separately on low and high scoring essays for essays written by speakers of each of 35 different native languages. Finally, we compared the distances between words across language pairs (e.g., correlation of distances for content words based on essays written by native Spanish versus native French speakers). x-axis shows magnitude of cross-linguistic word distance correlation (Pearson’s r); y-axis shows counts of language pairs. Facets show results from models trained on low (score $<$ 4 on 5 pt. scale; bottom) and high (score $\geq$ 4; top) scoring essays. Color indicates word type. Vertical lines indicate distribution means with bootstrapped 95% confidence intervals. **B**. Model results predicting the difference between local and global word correlations by language pair (N = 595). Intercept parameter compares local-global difference to zero; Grammatical distance predictor measures typological distance between languages based on the WALS database (16, 17). Top table shows results for Wikipedia corpus; Bottom table shows results for TOEFL corpus. These models suggest that languages are more similar locally, relative to globally, and that this difference is not predicted by grammatical similarity between languages.

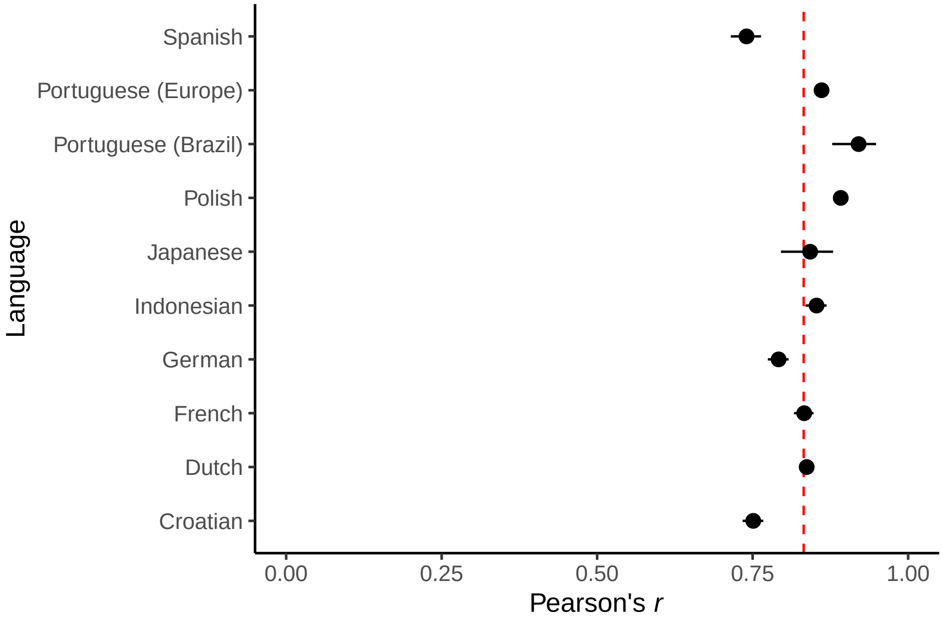


Fig. S5. Comparison of concreteness word norms collected in English versus estimates in other languages. We aggregated all available concreteness estimates collected by native speakers in a language other than English, and compared the estimate in each language to the estimate of its translation equivalent in English (18). Concreteness norms were available for 10 languages: Spanish (19), Portuguese (European) (20), Portuguese (Brazil) (21), Polish, (22) Japanese (23), Indonesian (24), German (25), French (26), Dutch (27), and Croatian (28). The figure shows the correlation (Pearson's r) between word estimates in each language, compared to English. Ranges correspond to 95% confidence intervals, and the red dashed line indicates the mean correlation across languages. Overall, estimates of concreteness collected in English were highly correlated with estimates collected in other languages (M = .83), suggesting a word's conceptual concreteness is largely language independent.

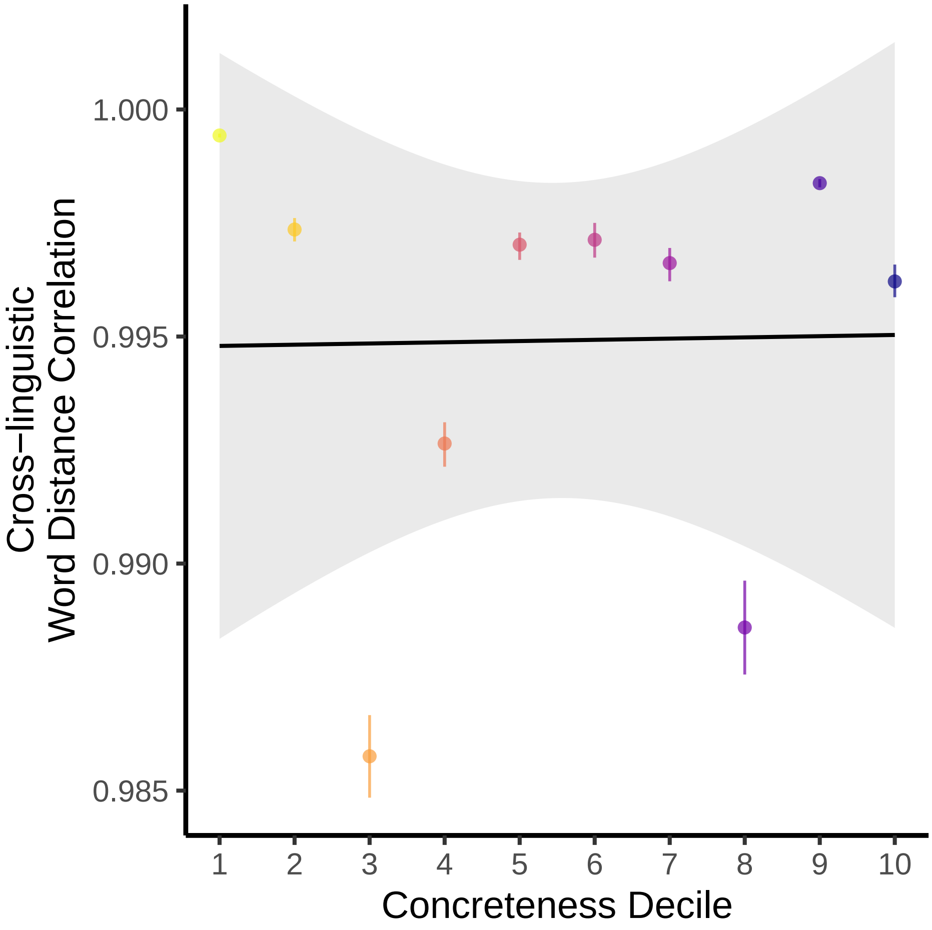


Fig. S6. An alternative explanation for the finding that concrete words are more similar cross-linguistically relative to abstract words is that concrete words, in general, tend to appear in more similar contexts in corpora of text relative to abstract words. To test this possibility, we constructed 35 corpora with each corpus containing equally-sized samples of TOEFL essays from each of the 35 different languages. Each corpus was therefore comparably sized to the language-based corpora described in the Main Text (approx. 1100 essays), but contained essays written by speakers of all 35 native languages. We then calculated mean pairwise correlation between word distances across languages as a function of the concreteness decile of the words. Point ranges correspond to bootstrapped 95% confidence intervals; range on model fit corresponds to the standard error. Unlike for the language-based corpus, there is no effect of concreteness on cross-corpus similarity: Across corpora, low and high concrete meanings are equally similar (r = .02; p = .96). This suggests that the concreteness effect reported in the Main Text is due to cross-linguistic differences in meanings, rather than an artifact of, e.g., differing types of linguistic contexts for high versus low concrete words.

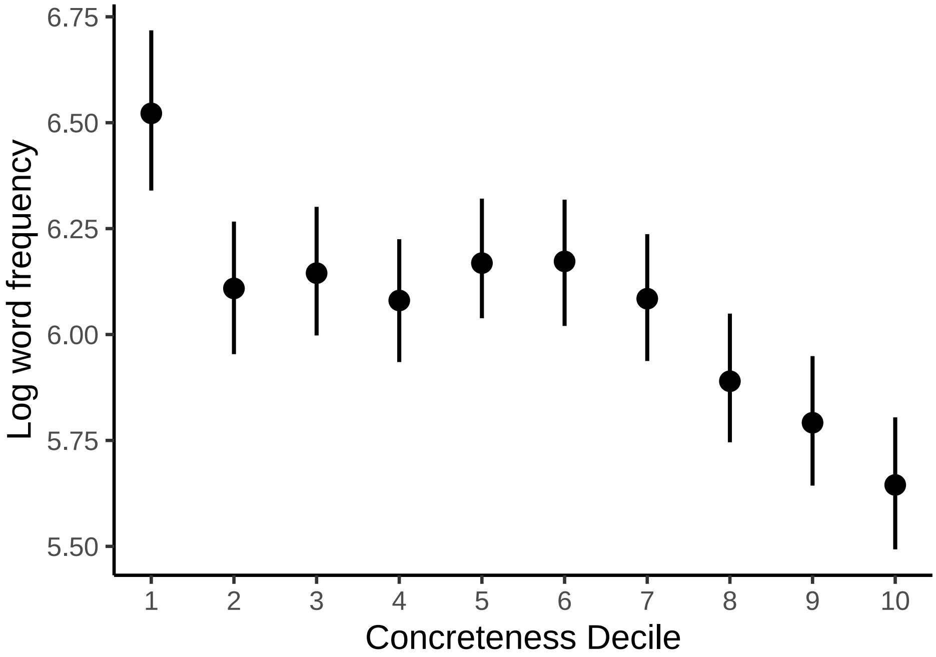


Fig. S7. Mean word frequency for words in each concreteness decile from the TOEFL corpus. Word frequency is estimated from the actual word counts in the corpus. Error bars are bootstrapped 95% confidence intervals. This analysis suggests that more concrete words tend to be less frequent in the corpus, making it unlikely that the relationship between concreteness and cross-linguistic similarity is due to measurement noise.

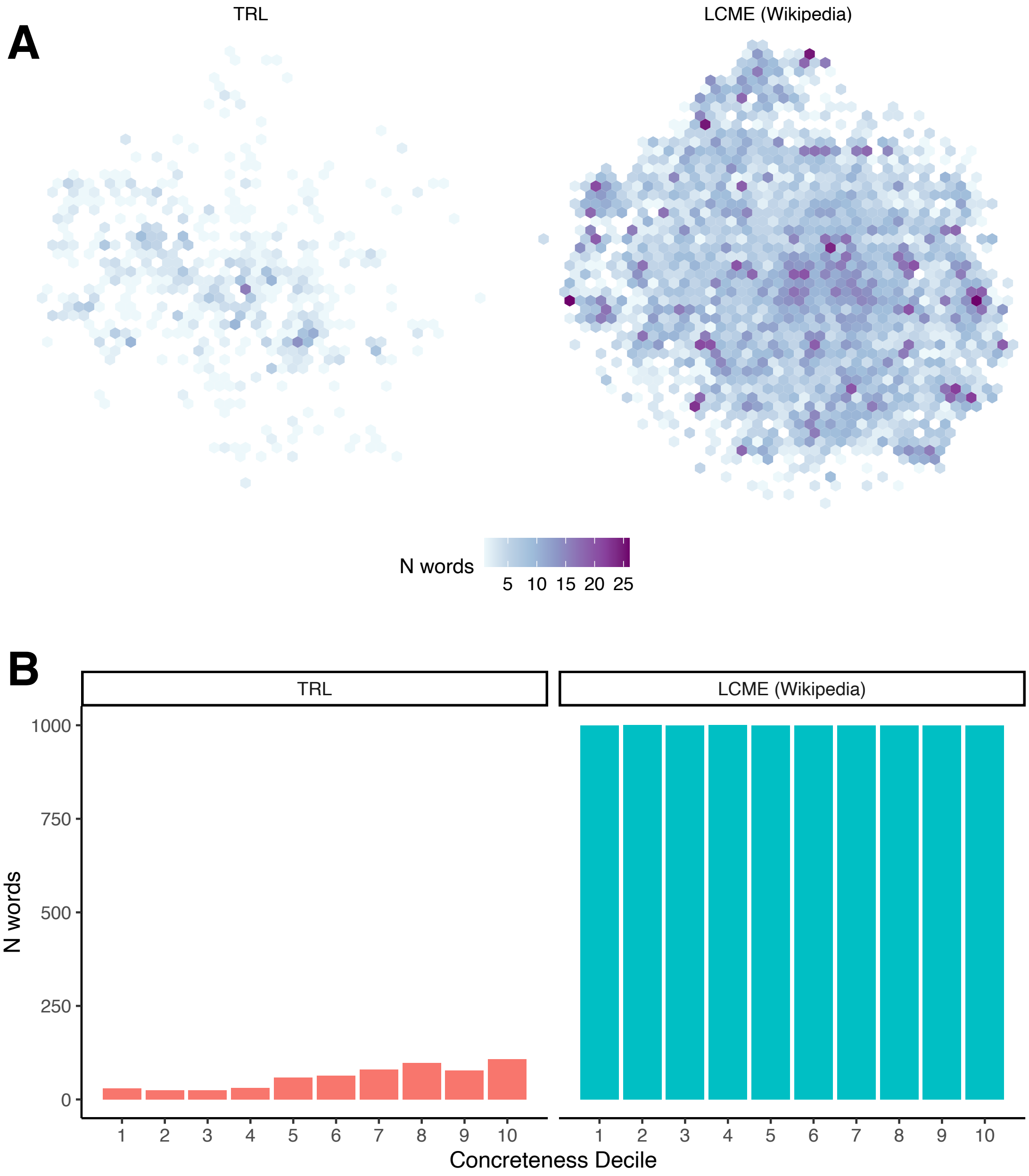


Fig. S8. Comparison of our samples (LCME) with those in (29), or TRL hereafter. **A**. Density plot covering our word samples from the English corpus, projected to two dimensions using TSNE; **B**. Histograms of the samples over the deciles of the concreteness distribution. TRL sampled concepts/words by 21 hand-selected domains (e.g., animals, food & drink, possessions, kinship, quantities, time) with the expressed purpose of exploring the relationship between the local coherence and category type across their sample of 10 language families (41 languages). They did this by selecting the 100 closest neighbors to each target word, and correlating these ranked lists across languages, then averaging them across target words within domain (e.g., "mother", "father", "aunt", "uncle" for the kinship domain). This local measurement revealed that the abstract domains of quantity, time and kinship correlated most highly across languages because of their internal coherence. They then inferred that abstract concepts in general were more coherent than concrete concepts. Plots A and B show that TRL's words unwittingly clumped in semantic space and within the higher deciles of the concreteness distribution – they were not selected to test the concreteness hypothesis. By contrast, we use 10,000 concepts randomly sampled from each decile of the concreteness norm distribution for the purpose of testing the relationship between concreteness/abstractness and variance in meaning across languages. Our samples show that the concrete-abstract association does correlate with word variation across languages, but explains less than word position. Words within domains (independent of concreteness) cluster together across languages much more tightly than between them, creating local coherence but global variance across the world's language.

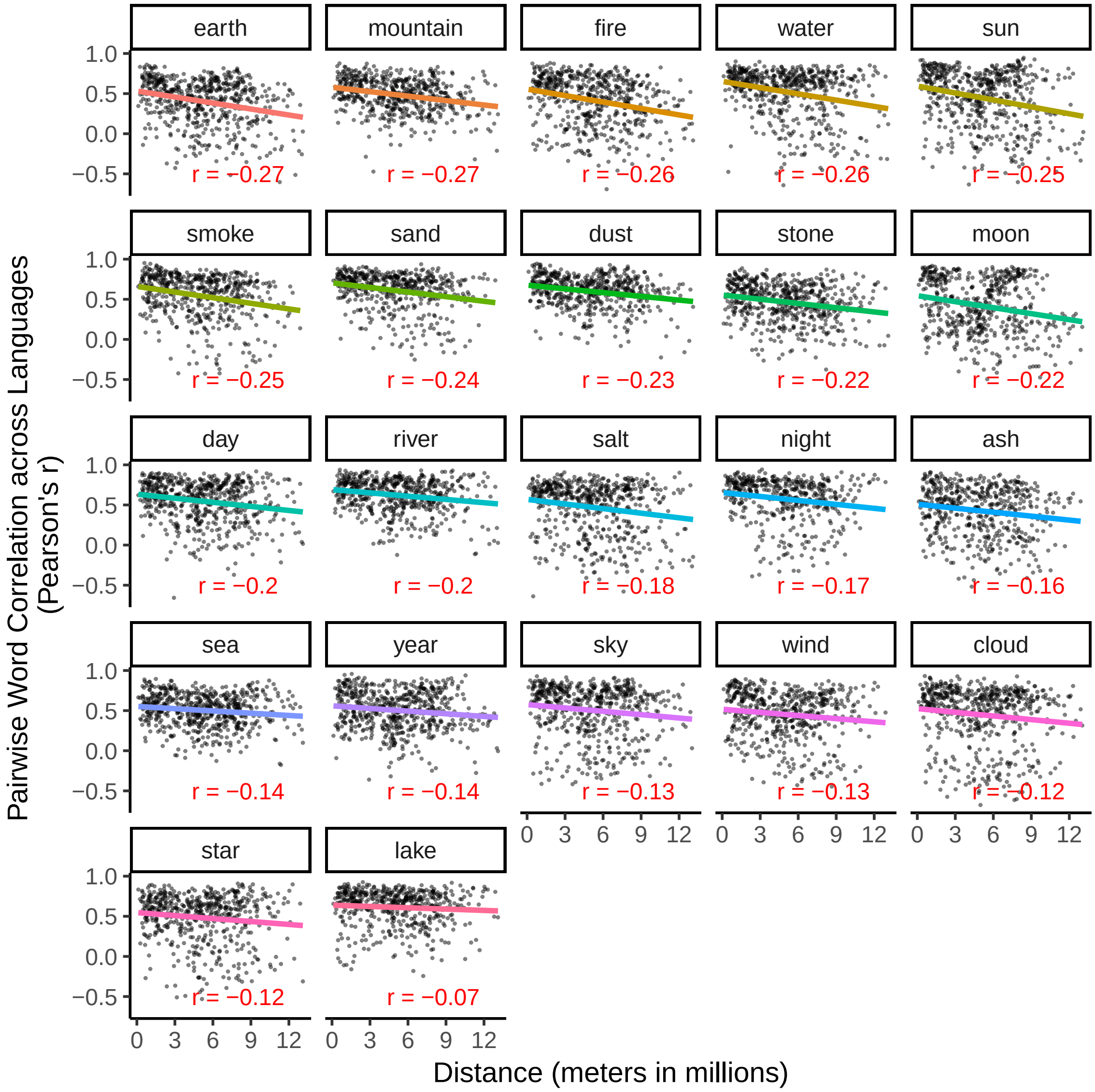


Fig. S9. Physical distance between where two languages are spoken predicts semantic similarity of Swadesh (30) meanings between two languages: Languages that are geographically closer have more similar meanings. Each facet corresponds to a Swadesh word. Each point shows the correlation between the pairwise distances between the target Swadesh word and all other Swadesh words for a language pair (e.g., a point on the "earth" facet represents the magnitude of the correlation between Spanish and French of the distances between "earth" and all other Swadesh words). Physical distance in meters (millions) between languages is plotted on the x-axis and magnitude of correlation between distances to other Swadesh items is plotted on the y-axis. All correlations are significant at the α = .01 level except the items "star" (QAP p = .016) and "lake" (QAP p = .119).

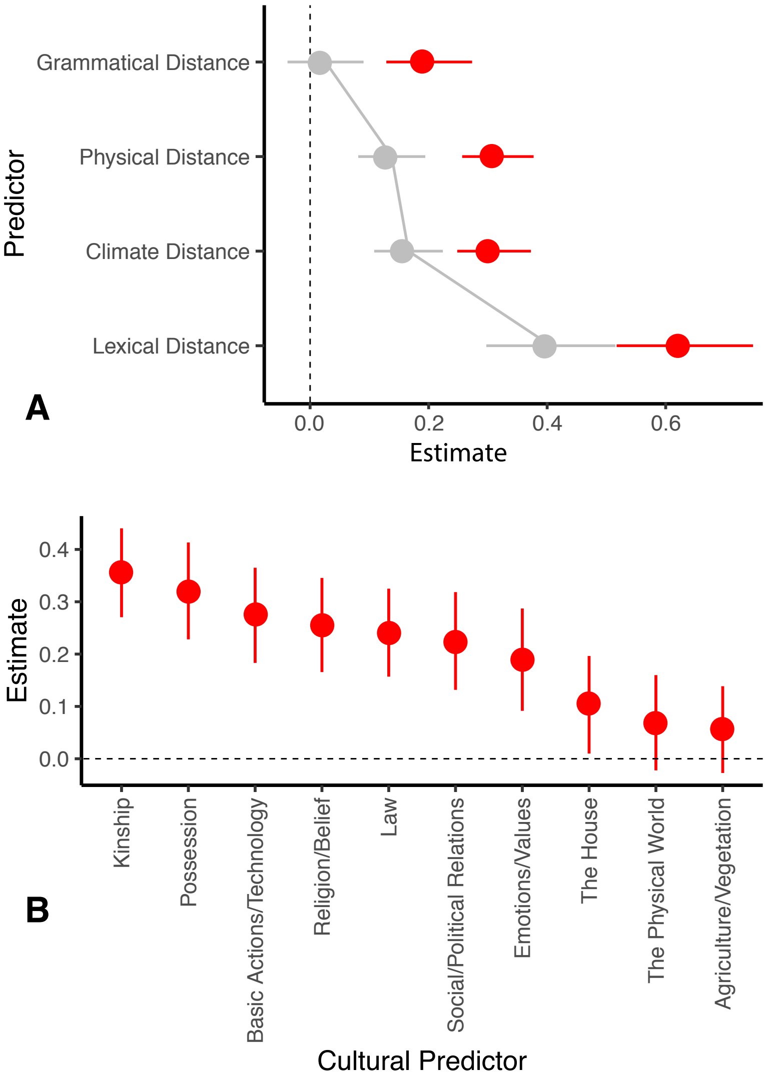


Fig. S10. A. Language pairwise predictors of semantic similarity for full sample of 35 languages. Cultural language distances were not available for 7 of the 35 languages in our sample (German, Greek, Italian, Portuguese, Romanian, Farsi, and Yoruba). In the Main Text, we report model parameters for the set of 28 languages for which we have full data. Here, we show model parameters for the other predictors for the full sample of 35 languages. Red points indicate standardized estimates from a single-predictor model; grey points indicate estimates from additive linear model with all five predictors included. Ranges are 95% confidence intervals. **B**. Cultural predictors of semantic distance for each of 10 cultural sub-domains. Points are standardized estimates from an additive linear model with all five predictors included. Ranges are 95% confidence intervals.

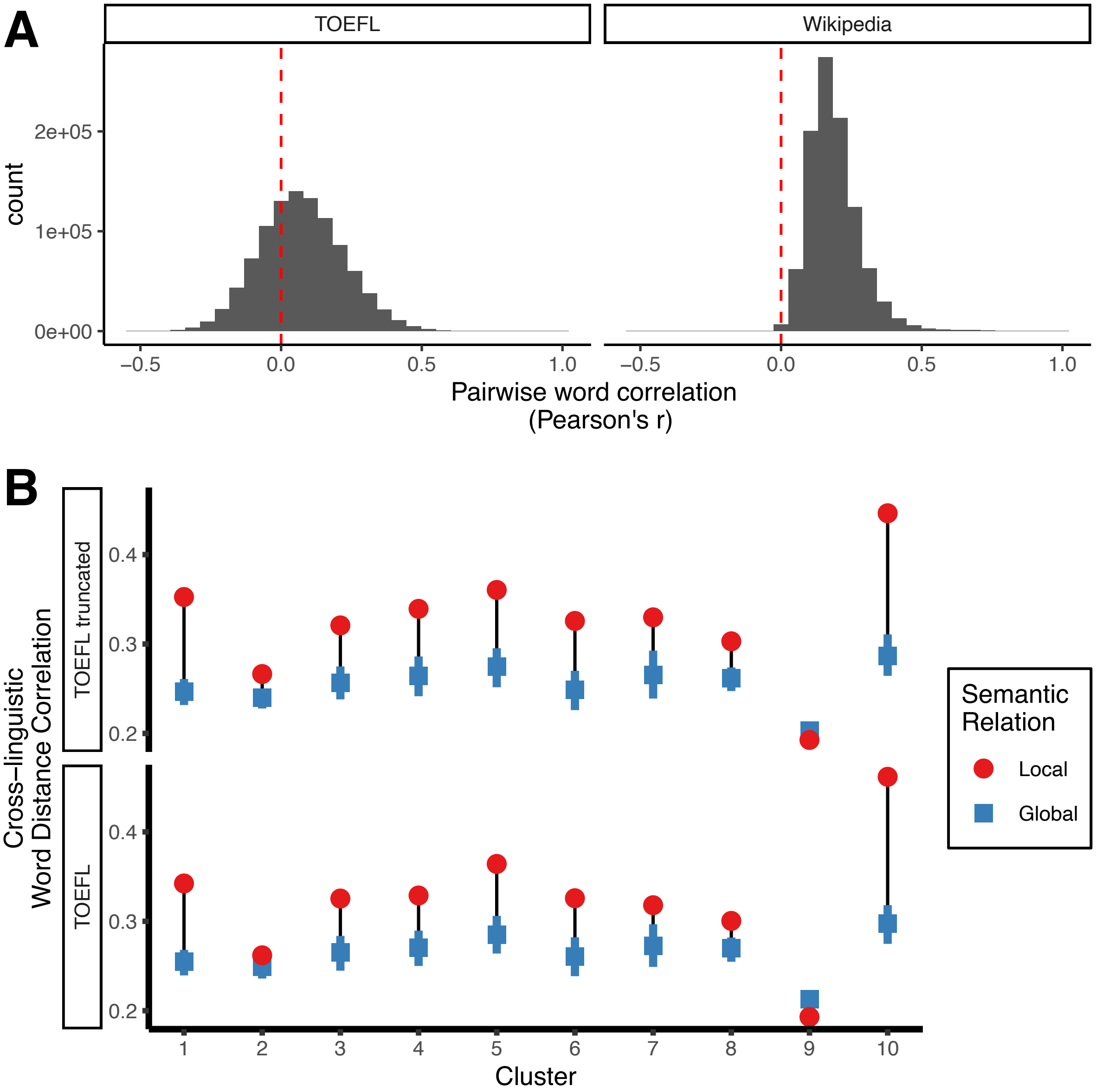


Fig. S11. Local vs. global distances, examining (A) then truncating (B) negative distances. **A**, the distribution of pairwise distances within language models from a random sample of 1M word pairs in the TOEFL models (left) and Wikipedia models (right). For Wikipedia, 99.8% of pairwise distances were greater than zero. However, in the case of the TOEFL models, only 67.9% were greater than zero, meaning there were a large subset of large distances included in our analyses (32.1%). **B**, Because very large distances (e.g., those larger than 90° with a cosine $\theta < 0$) are unlikely to be meaningfully different than those defined as unrelated and equal to 90° and $\theta = 0$ because they were not trained with inverse associations, we re-ran our core analysis (see Fig. 3C), rounding cosine distances greater than 90° (i.e., negative values) to 90° – the theoretical bound for unrelatedness. The result represents a nearly 20% increase in the local-global effect (Truncated TOEFL: M = 0.069, SD = 0.008; t(594) = 205.24; p < .0001; d = 3.4 [3.22, 3.58]; W = 177310, p < .0001; TOEFL: M = 0.058, SD = 0.008; t(594) = 185.97; p < .0001; d = 2.84 [2.68, 3]; W = 177310, p < .0001). This suggests that the analyses reported in the main text, which included these very large distances, lead to a conservative estimate of the local-global effect for the TOEFL data.

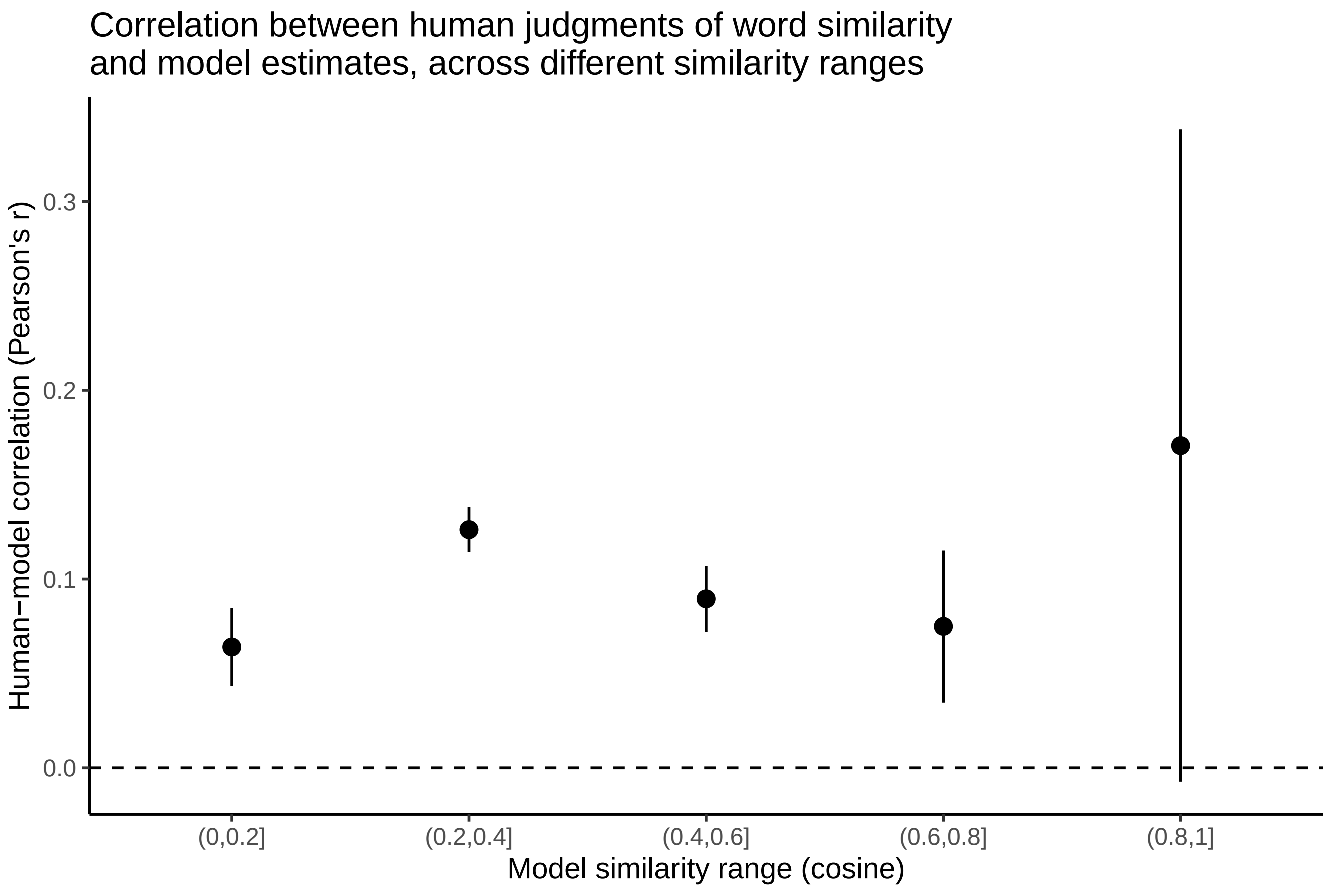


Fig. S12. Stable correlations between close and distant words in our sample. Data from word-pairs crowd-sourced by participants in the Small World of Words project – https://smallworldofwords.org/en/project/ (31) – in which participants were given a target word (e.g., “cat”) and asked to list three associated words (e.g., “kitten”, “milk”, “mouse”). We analyzed the first associates for each target word, which are likely to be most closely associated with the target. By aggregating across many participants, we can estimate the similarity between the target and associate as the conditional probability of the associated word, given the target (P(“kitten”|“cat”). We then get an estimate of model similarity by projecting target and associated words onto a FastText model built from English Wikipedia (32), and calculating their cosine distance. The plot shows the correlations between human and model estimates for word pairs at varying cosine distances. As a group, words farther from the focal word (i.e., more distal words) are comparably correlated to human judgments than words much closer. To the extent that these models capture psychological similarity between word meanings, this means that they do so equally well for words close and farther in meaning from the target word.

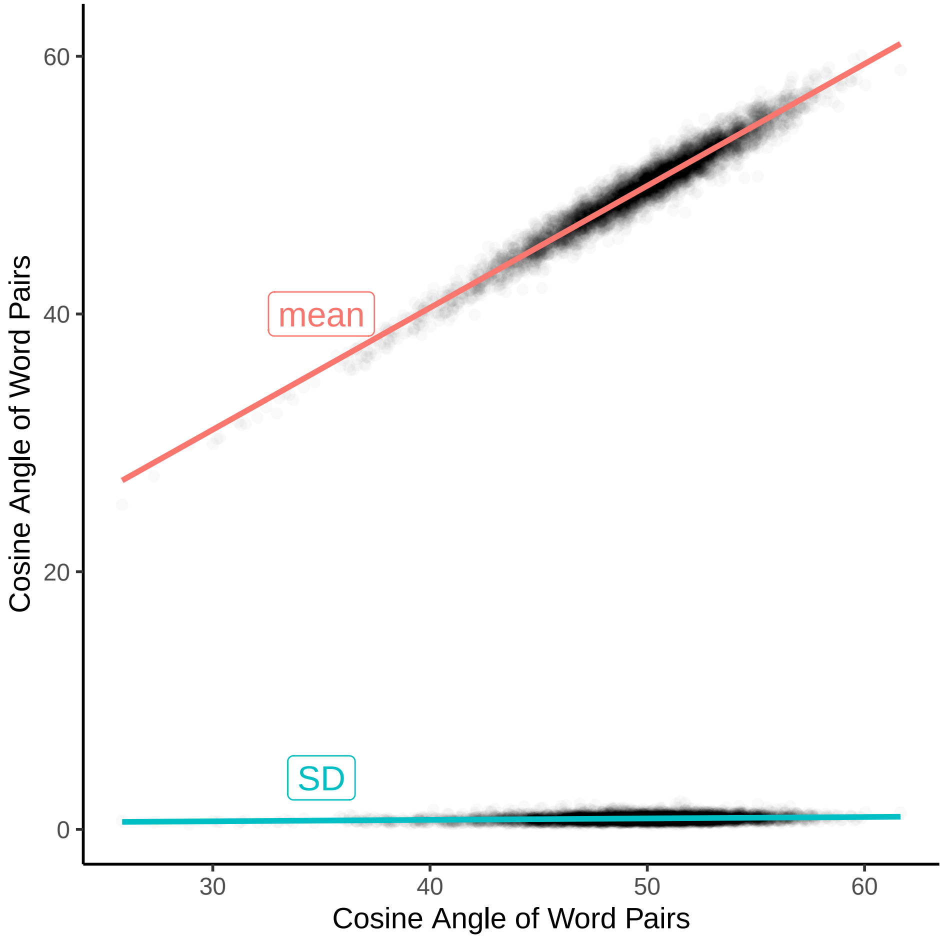


Fig. S13. Analysis of the relationship between mean and standard deviation for the cosine angle between pairs of word vectors from word embeddings built from a corpus of Google Books n-grams (2000-2010). Means and standard deviations for 100 words randomly sampled from our ETS analysis, computed from 20 word2vec models built from nonoverlapping subsets of the corpus (33) – details in (7) – resulting in 4,900 unique word pairs between distinct words. Greater mean cosine distances (measured in degrees) between words enables a much greater theoretical variance and standard deviation between those distances across models. Nevertheless, our subsampled standard deviations barely increase at all. This strongly suggests that distant associations are at least as stable and precise across models (and the distinct subsamples of text on which they are trained) as proximate associations. Distant associations cannot be attributable to random seeds or fluctuations of the algorithm.

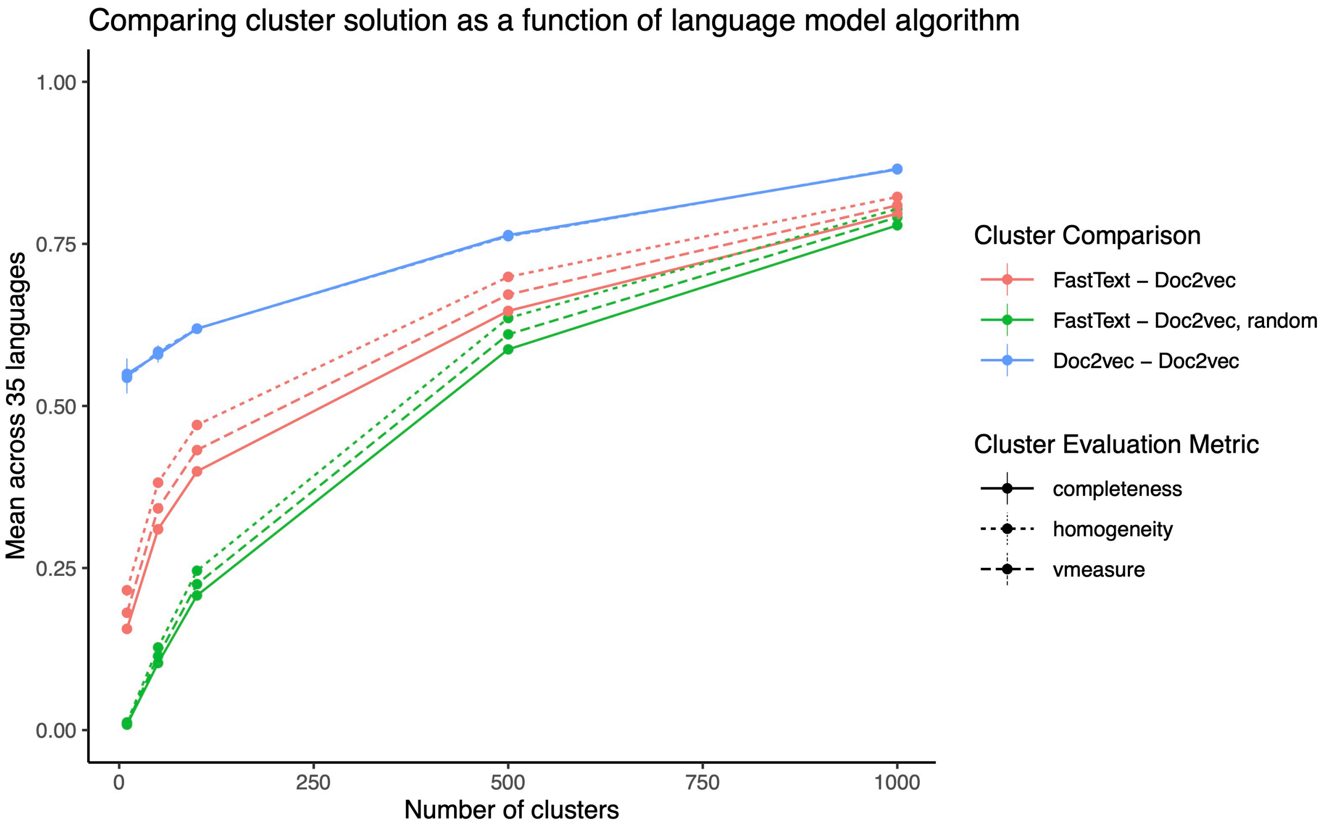


Fig. S14. Comparison of FastText and Doc2Vec algorithms. We directly compare FastText and Doc2Vec algorithms by training new models with each on the ETS essay data. For both models, the distance between the same word pairs correlate substantially and significantly. We analyzed the similarity of clusters constructed from FastText model outputs with those constructed from Doc2Vec model output using the V-measure (34), which 1) does not depend on the clustering algorithm or data set, 2) evaluates the clustering of all data points, and 3) constructs an accurate evaluation and combination (the geometric mean) of homogeneity and completeness. For example, a FastText clustering result satisfies homogeneity if all of its clusters contain only data points that are members of a single Doc2Vec cluster. A clustering result satisfies completeness if all the data points that are members of a given FastText cluster are elements of the same Doc2Vec cluster. In this way, the homogeneity and completeness of a clustering solution run approximately in opposition: Increasing homogeneity often results to decrease its completeness. We compare 10, 50, 100, 500, and 1000 cluster solutions from FastText and Doc2Vec with respect to each other and upper and lower baselines: (1) Doc2Vec vs. Doc2Vec clusters, constructed with different random seeds, which represents an upper bound, and (2) Doc2Vec vs. random clusters, which represents a lower bound. For the 10 cluster solutions, mean V-measures are approximately .2, rising to nearly .4 in the 100 cluster solution. For all but the largest cluster solutions (1000), these are much higher than the random baseline. Note that as we increase the cluster solution from 10 to 100 clusters, we increase the similarity between the FastText and Doc2Vec clusterings relative to the upper baseline and decrease its similarity to the lower one.

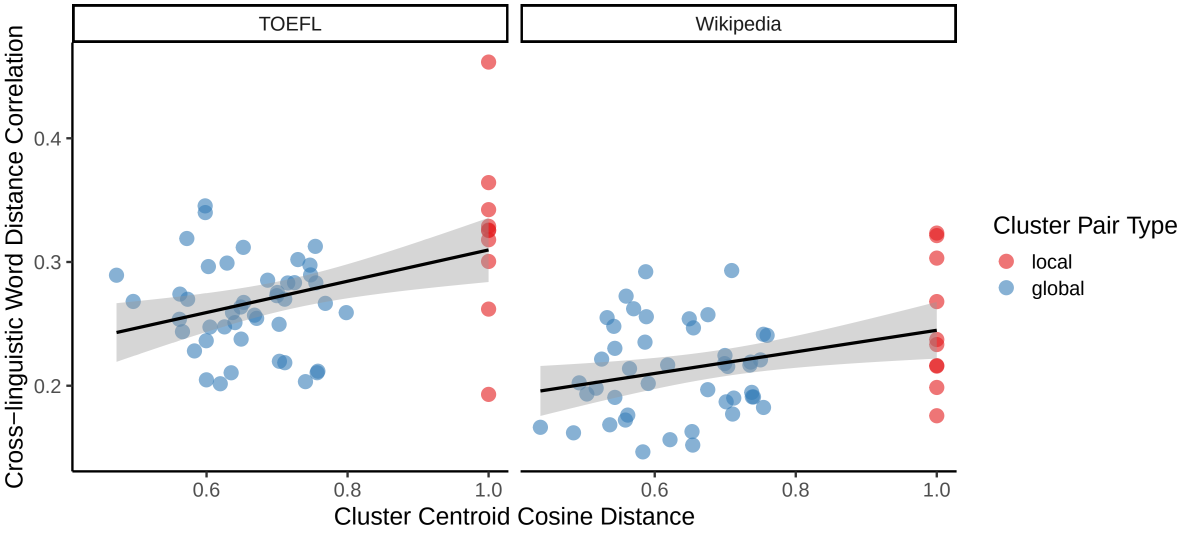


Fig. S15. As described in the Main Text, we clustered words into 10 clusters for each of the two corpora (Second-Language TOEFL and Multilingual Wikipedia). Each point in the figure corresponds to a unique pair of clusters. The y-axis shows the average correlation (Pearson's r) in pairwise word distances between the two clusters across all language pairs (e.g., correlation of all word pair distances between the 1-2 cluster pair for French and Spanish, and all other language pairs). The x-axis shows the cosine distance between the centroids of the two clusters. Color indicates whether the two clusters are the same ("local;" red) or different ("global;" blue). There is a positive correlation between cluster centroid distance and the magnitude of the pairwise word correlation for both corpora (TOEFL: r = .39, p = .003; Wikipedia: r = .34, p = .01). Notably, however, there is a bimodal relationship in centroid distances characterized by the local-global distinction.

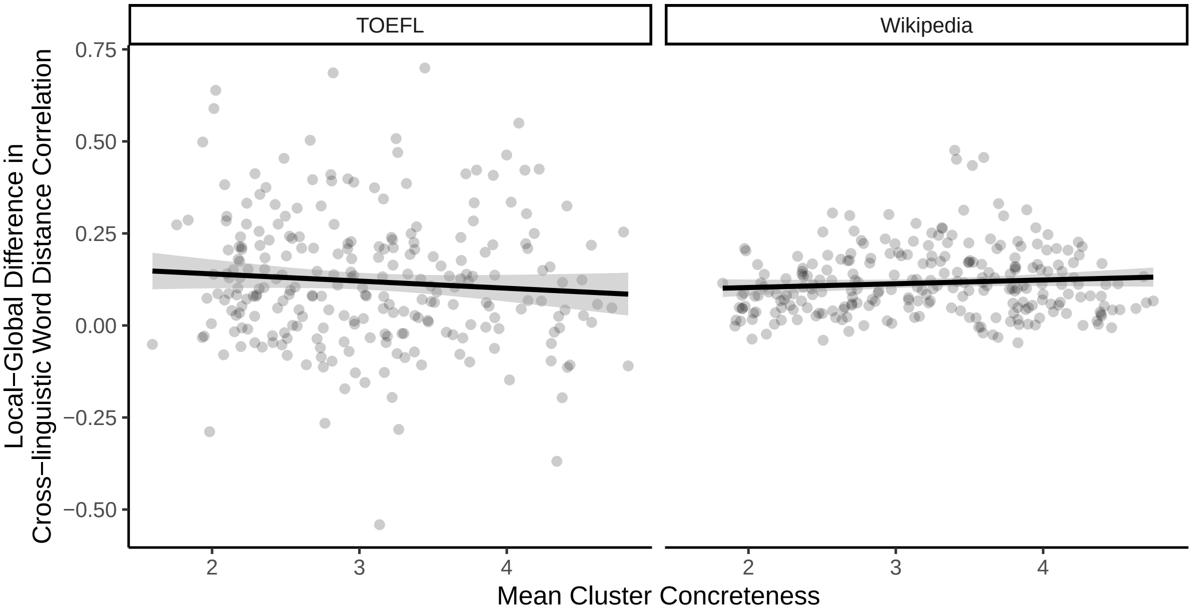


Fig. S16. Relationship between cluster concreteness and the magnitude of the local-global effect for the 250 cluster solution for TOEFL (left) and Wikipedia (right). For each cluster in each corpus, we calculated the mean concreteness of words in that cluster and the mean cross-linguistic correlation in word pairwise distances within the same cluster (local) versus across different clusters (global). The y-axis shows the difference in local versus global correlations aggregating across language pairs. Larger values indicate that word meanings are more similar within clusters versus across them. Each point corresponds to a cluster. These results reveal that, while semantic clusters tend to co-vary with concreteness, the concreteness of a cluster is not predictive of the local-global effect.

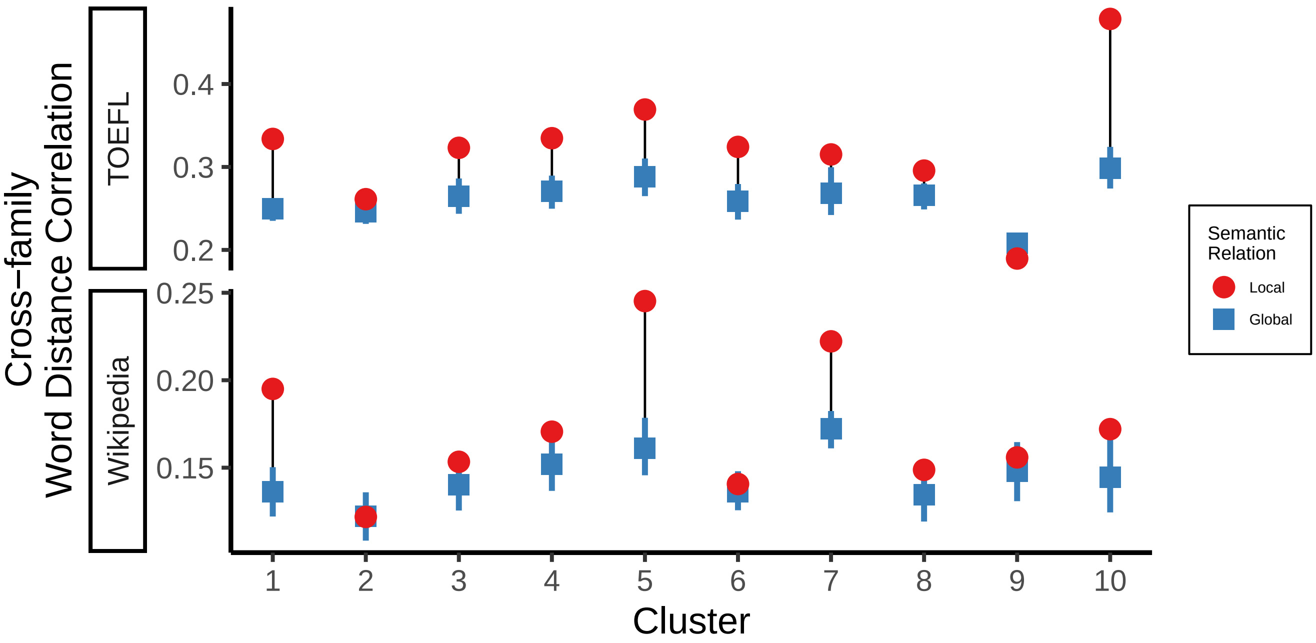


Fig. S17. Cross-family (11 language families) word distance correlations for word groups in 10 semantic clusters based on word embeddings obtained from Second-Language TOEFL Corpus (top) and Multilingual Wikipedia Corpus (bottom). Red points indicate mean local correlation for each cluster, and blue squares indicate global correlation for each cluster (TOEFL: $M = 0.061$, $SD = 0.007$; $t(54) = 64.37$; $p < .0001$; $d = 2.98$ [2.44, 3.52]; Wikipedia: $M = 0.03$, $SD = 0.033$; $t(54) = 6.91$; p < .0001; d = 0.29 [-0.09, 0.66]).Note that the cluster number labels are arbitrarily assigned, and that the words assigned to each cluster differ across the two corpora. Despite far fewer observations than in the main analysis, we see the same, statistically significant pattern in both corpora: local distance are more highly correlated across language families, relative to global distances. Even though several linguistic families manifest a single language representative (e.g., Afro-Asiatic, Altaic, Austro-Asiatic, Japonic, Sino-Tibetan, Tai-Kadai, Korean), with necessarily smaller Wikipedia text corpora, the pattern is manifest even with these more limited textual samples.

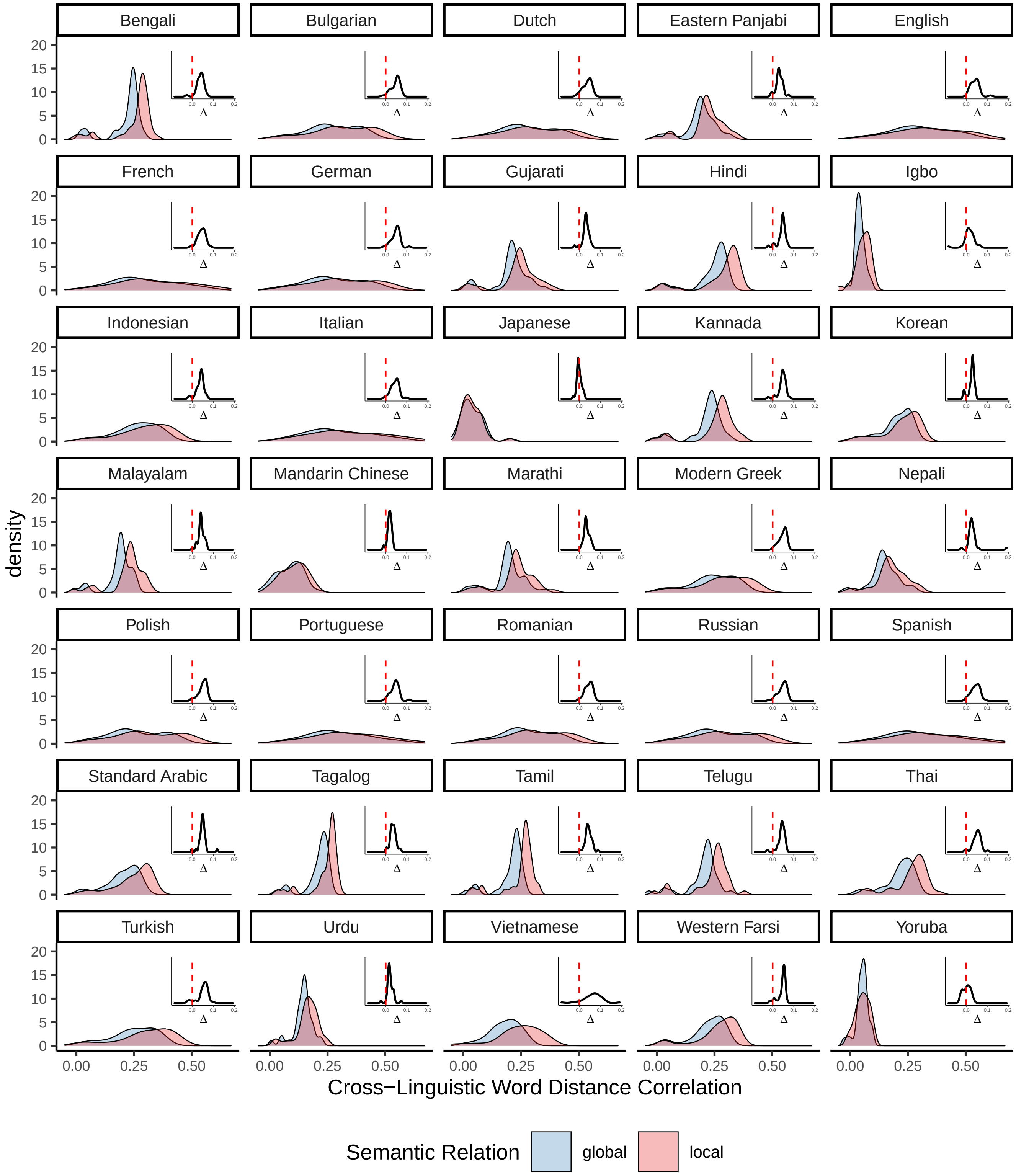


Fig. S18. Distribution of local-global correlation difference by language, for the Wikipedia corpus. Each subplot shows the difference in correlation between local distances and global distances for one language with all other languages (n = 34). The insets manifest the differences in the two distributions with the dashed red line signifying a difference of 0. These demonstrate that Wikipedia local-global differences are not driven by outliers but appear largely balanced across all language combinations, with a consistent effect of roughly the same size across language comparisons.

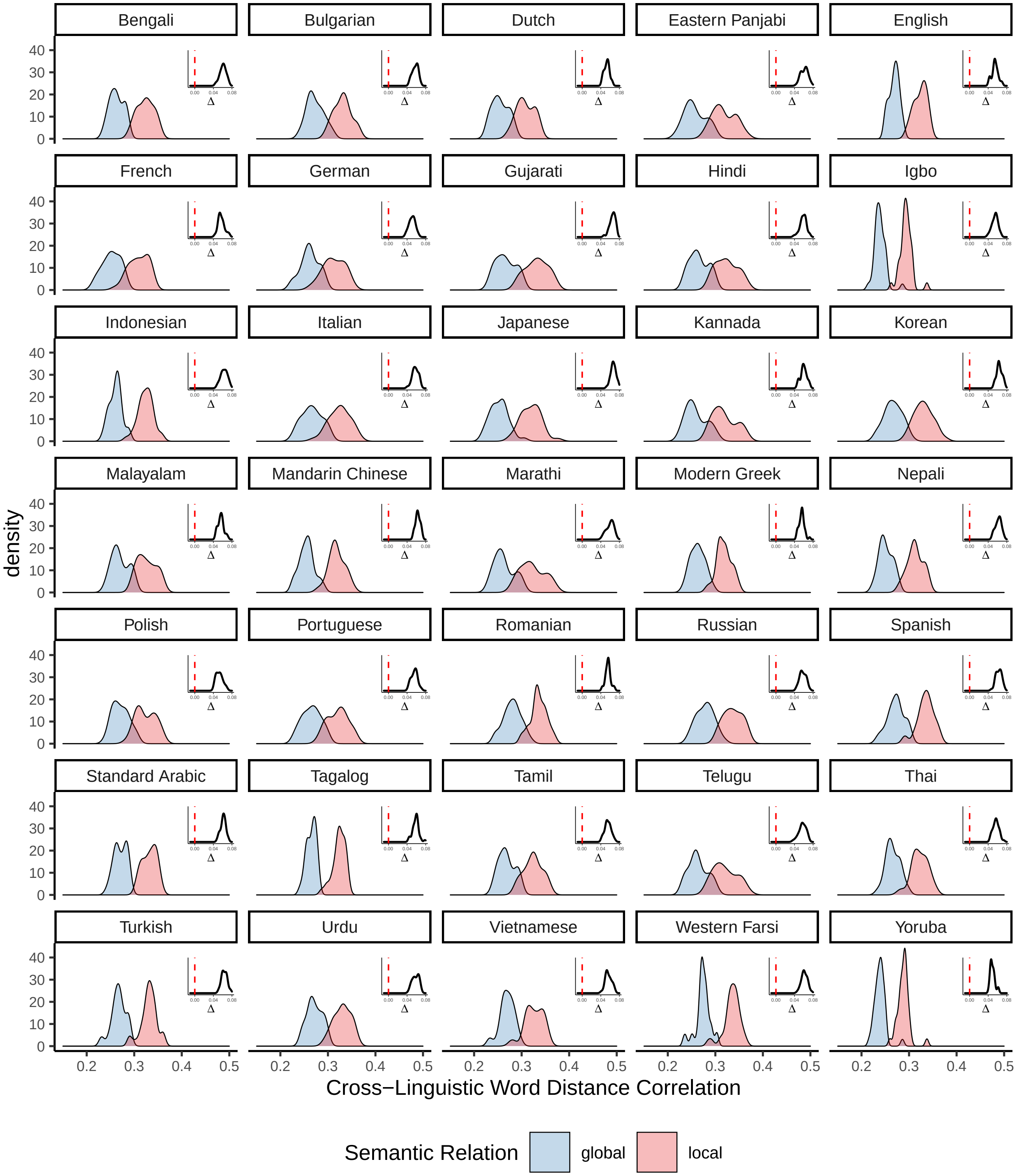


Fig. S19. Distribution of local-global correlation difference by language, for TOEFL corpus. Each subplot shows the difference in correlation between local distances and global distances for one language with all other languages (n = 34). The insets manifest the differences in the two distributions with the dashed red line signifying a difference of 0. These demonstrate that TOEFL local-global differences are not driven by outliers but appear largely balanced across all language combinations, with a consistent effect of roughly the same size across language comparisons.

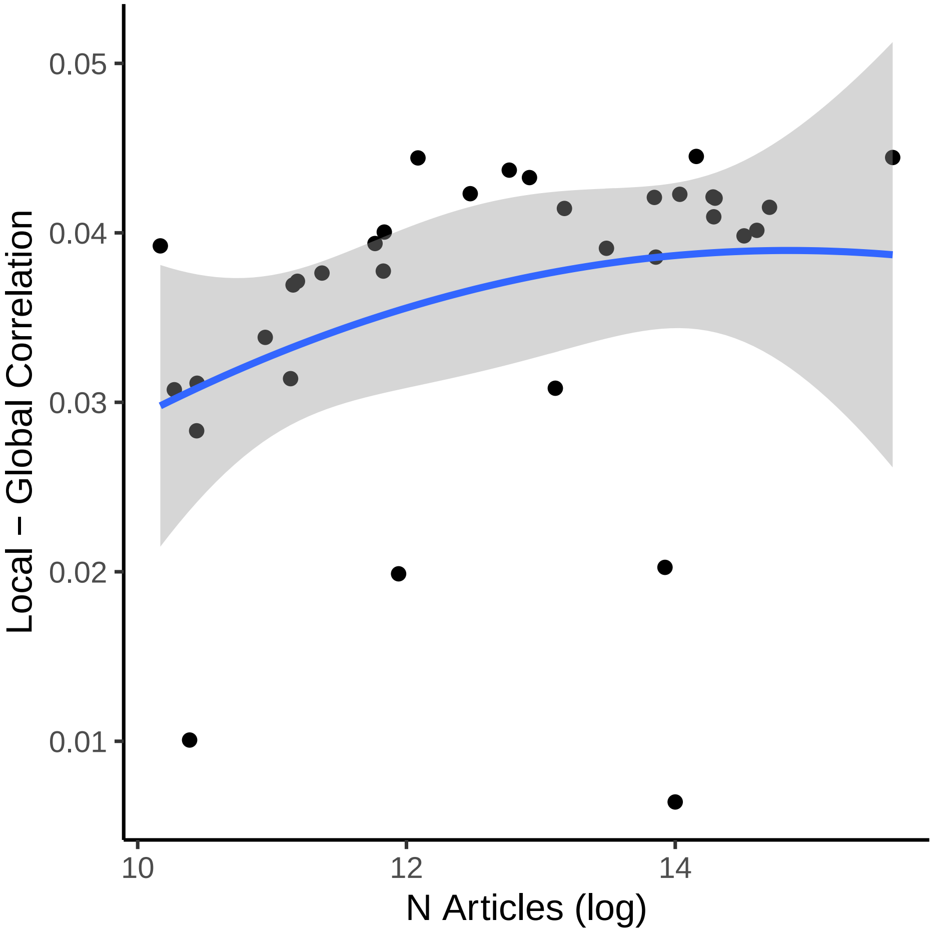


Fig. S20. Difference in cross-linguistic word distance correlations between local and global relations, as a function of the number of Wikipedia articles present for a given language (1). Each point corresponds to a language. Cross-linguistic word distance correlations are estimated from models trained on Wikipedia in each language. While in all languages local correlations are greater than global correlations, the difference is somewhat larger for languages for which more articles are available. This is almost certainly due to measurement error for local words. Smaller collections of articles will necessarily have smaller numbers of local (within-cluster) words, leading to higher variance and lower local-word correlations between languages. These lower localized correlations reduce the difference between local and global semantic distances. Nevertheless, when we split the languages in half by Wikipedia page number, we find that the significance of the relation holds for both small and large Wikipedia page languages.

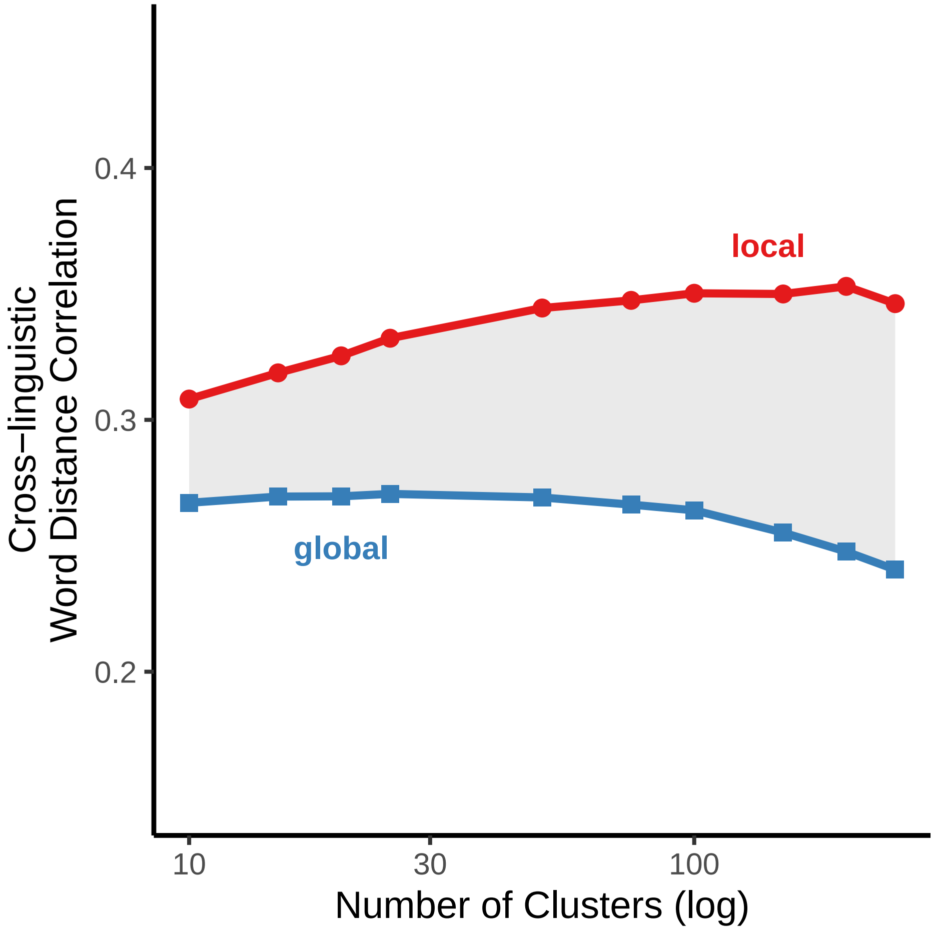


Fig. S21. Cross-linguistic word distance correlations for local (red) versus global (blue) semantic comparison as a function of the number of semantic clusters using native model clusters for models trained on TOEFL essays. In Fig. 3D of the Main Text, we demonstrate that distances within clusters (local) are more correlated with each other cross-linguistically than word distances across clusters (global). This analysis was based on clusters defined by a model trained on English Wikipedia articles. Here, we replicate this pattern using clusters determined by models trained on native language text. For example, when comparing TOEFL essays from native Hindi and Mandarin speakers, we clustered words covering the Hindi and Chinese Wikipedia entries to capture how each language represents its knowledge base. Then we compared within Hindi-Wikipedia-clusters vs. between Hindi-Wikipedia-clusters for essays from native speakers of both languages; next we compared within Mandarin-Wikipedia-clusters vs. between Mandarin-Wikipedia clusters for the same essays; finally we averaged these differences.

		Semantic	Grammatical	Physical	Climate	Lexical	Cultural
35 languages	Semantic	1.00					
	Grammatical	0.55	1.00				
	Physical	0.52	0.46	1.00			
	Climate	0.55	0.35	0.55	1.00		
	Lexical	0.71	0.60	0.45	0.45	1.00	
28 languages	Semantic	1.00					
	Grammatical	0.61	1.00				
	Physical	0.52	0.53	1.00			
	Climate	0.54	0.41	0.58	1.00		
	Lexical	0.73	0.63	0.47	0.49	1.00	
	Cultural	0.69	0.47	0.49	0.59	0.72	1.00

Table S1. *Top:* **Pairwise correlations (Pearson's r) between language distances measures for all 35 languages (excluding cultural distance measure which has missing data).** *Bottom:* **Pairwise correlations (Pearson's r) between language distance measures for the subset of 28 languages for which data is available for all 6 predictors. All correlations are significant at the α = .05 level using QAP tests.**

References

1. List of Wikipedias by speakers per article (2019) [Online; accessed 25-May-2020].
2. T Amano, et al., Global distribution and drivers of language extinction risk. *Proc. Royal Soc. B: Biol. Sci.* **281**, 20141574 (2014).
3. Productivity of Wikipedia authors (2012) [Online; accessed 25-May-2020].
4. Analytics/archive/data/pagecounts-raw — Wikitech (2019) [Online; accessed 25-May-2020].
5. SL Wu, Influence of L1 thinking for speaking on use of an L2: The case of path expressions by English-Speaking learners of Chinese. *Stud. Second. Lang. Acquis. Chin.*, 1–29 (2014).
6. L Filipović, Speaking in a second language but thinking in the first language: Language-specific effects on memory for causation events in English and Spanish. *Int. J. Biling.* **22**, 180–198 (2018).
7. AC Kozlowski, M Taddy, JA Evans, The geometry of culture: Analyzing the meanings of class through word embeddings. *Am. Sociol. Rev.* **84**, 905–949 (2019).
8. MH Bodell, M Arvidsson, M Magnusson, Interpretable word embeddings via informative priors. *arXiv preprint arXiv:1909.01459* (2019).
9. J An, H Kwak, YY Ahn, Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. *arXiv preprint arXiv:1806.05521* (2018).
10. H Kwak, J An, YY Ahn, FrameAxis: Characterizing framing bias and intensity with word embedding. *arXiv preprint arXiv:2002.08608* (2020).
11. O Levy, Y Goldberg, Linguistic regularities in sparse and explicit word representations in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning.* pp. 171–180 (2014).
12. O Levy, Y Goldberg, Neural word embedding as implicit matrix factorization in *Advances in Neural Information Processing Systems.* pp. 2177–2185 (2014).
13. O Levy, Y Goldberg, I Dagan, Improving distributional similarity with lessons learned from word embeddings. *Transactions Assoc. for Comput. Linguist.* **3**, 211–225 (2015).
14. R Jakobson, *On language.* (Harvard University Press), (1990).
15. M Brysbaert, B New, E Keuleers, Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behav. Res. Methods* **44**, 991–997 (2012).
16. D Dediu, Making genealogical language classifications available for phylogenetic analysis: Newick trees, unified identifiers, and branch length. *Lang. Dyn. Chang.* **8**, 1 – 21 (2018).
17. MS Dryer, M Haspelmath, eds., *WALS Online.* (Max Planck Institute for Evolutionary Anthropology, Leipzig), (2013) http://wals.info.
18. M Brysbaert, AB Warriner, V Kuperman, Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **46**, 904–911 (2014).
19. M Guasch, P Ferré, I Fraga, Spanish norms for affective and lexico-semantic variables for 1,400 words. *Behav. Res. Methods* **48**, 1358–1369 (2016).
20. AP Soares, AS Costa, J Machado, M Comesaña, HM Oliveira, The Minho word pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. *Behav. Res. Methods* **49**, 1065–1081 (2017).
21. MFRd Lima, LG Buratto, Norms for familiarity, concreteness, valence, arousal, wordlikeness, and recall accuracy for Swahili–Portuguese word pairs. *SAGE Open* **11**, 2158244020988524 (2021).
22. KK Imbir, Affective norms for 4900 Polish words reload (anpw_r): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. *Front. Psychol.* **7**, 1081 (2016).
23. D Allen, K Conklin, Cross-linguistic similarity norms for Japanese–English translation equivalents. *Behav. Res. Methods* **46**, 540–563 (2014).
24. A Sianipar, P van Groenestijn, T Dijkstra, Affective meaning, concreteness, and subjective frequency norms for Indonesian words. *Front. psychology* **7**, 1907 (2016).
25. J Charbonnier, C Wartena, Predicting the concreteness of German words. in *SwissText/KONVENS.* (2020).
26. P Bonin, A Méot, A Bugaiska, Concreteness norms for 1,659 French words: Relationships with other psycholinguistic variables and word recognition times. *Behav. Res. Methods* **50**, 2366–2387 (2018).
27. M Brysbaert, M Stevens, S De Deyne, W Voorspoels, G Storms, Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychol.* **150**, 80–84 (2014).
28. B Ćoso, M Guasch, P Ferré, JA Hinojosa, Affective and concreteness norms for 3,022 Croatian words. *Q. J. Exp. Psychol.* **72**, 2302–2312 (2019).
29. B Thompson, SG Roberts, G Lupyan, Cultural influences on word meanings revealed through large-scale semantic alignment. *Nat. Hum. Behav.* **4**, 1029–1038 (2020).
30. M Swadesh, Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proc. Am. Philos. Soc.* **96**, 452–463 (1952).
31. S De Deyne, DJ Navarro, A Perfors, M Brysbaert, G Storms, The "small world of words" english word association norms for over 12,000 cue words. *Behav. research methods* **51**, 987–1006 (2019).
32. P Bojanowski, E Grave, A Joulin, T Mikolov, Enriching word vectors with subword information. (2016).
33. DN Politis, JP Romano, M Wolf, *Subsampling.* (Springer Science & Business Media), (1999).
34. A Rosenberg, J Hirschberg, V-measure: A conditional entropy-based external cluster evaluation measure in *Proceedings of*

the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). pp. 410–420 (2007).