





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

Molly Lewis^{a,1}, Aoife Cahill^b, Nitin Madnani^c , and James Evans^{d,e,1} 

Edited by Susan Fiske, Princeton University, Princeton, NJ; received January 17, 2023; accepted November 6, 2023

How does meaning vary across the world's languages? Scholars recognize the existence of substantial variability within specific domains, ranging from nature and color to kinship. The emergence of large language models enables a systems-level approach that directly characterizes this variability through comparison of word organization across semantic domains. Here, we show that meanings across languages manifest lower variability within semantic domains and greater variability between them, using models trained on both 1) large corpora of native language text comprising Wikipedia articles in 35 languages and also 2) Test of English as a Foreign Language (TOEFL) essays written by 38,500 speakers from the same native languages, which cluster into semantic domains. Concrete meanings vary less across languages than abstract meanings, but all vary with geographical, environmental, and cultural distance. By simultaneously examining local similarity and global difference, we harmonize these findings and provide a description of general principles that govern variability in semantic space across languages. In this way, the structure of a speaker's semantic space influences the comparisons cognitively salient to them, as shaped by their native language, and suggests that even successful bilingual communicators likely think with "semantic accents" driven by associations from their native language while writing English. 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, but because associations, metaphors, and narratives interlink meanings in different, predictable ways from one language to another.

human cognition | language | semantics | culture | communication

The degree to which word meanings vary across the world's languages and cultures is a fundamental question in the social and communication sciences. What precisely is the relationship between the meaning of "animal," "food," and "religion" in English and their closest translations in Persian, Hindi, and Russian? Variability in semantic structure is necessarily constrained by speakers' shared cognitive systems and the communicative functions demanded by social life (1–12). Nevertheless, there is now evidence for substantial variability across languages regarding their semantic organization (13–24). This evidence is primarily limited to the study of specific semantic domains, however, such as color (25), kinship (9, 26), and emotion (6).

In this paper, we describe and explore semantic variability by examining semantic relationships for many distinct languages across referential domains, rather than within a single domain. By taking this "systems-level" approach (27, 28), we seek to generalize prior work on meaning alignment to the macro structure of variability within and between semantic domains. Our results demonstrate variation within domains, with more concrete concepts translating more faithfully across languages than abstract ones. Moreover, we find substantially greater variation in meanings across domains. Languages manifest broad similarity in how they cluster words with meanings proximate to one another but diverge in how those clusters relate across semantic space. Across languages, meanings locally cohere but globally vary. For example, words associated with foods, body parts, spiritual agents, and human tragedies individually tend to cluster in similar ways, but relations between those clusters range widely by language.

Understanding the precise 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 (29). 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 (30). For example, when domains are close in a language's semantic space, the associations, analogies, metaphors, and narrative turns that interlink them may seem

Significance

The degree to which meanings align across the world's languages suggests the limits of translation and cross-cultural communication. Approaching this question demands a systems-level view to compare the structure of meanings across many languages. Using machine learning we construct word embeddings—dense, continuous, high-dimensional spaces that characterize word meanings from context—across large samples of multi-lingual text. With these representations, we find that 1) meanings across languages are similar within semantic domains and variable across them and that 2) concrete meanings are less variable across languages than abstranes, but all vary with distance. This suggests that associations and analogies, which interlink meanings within language, propose predictably different intuitions across distinct languages and confound the transmission of complex ideas.

Author contributions: M.L. and J.E. designed research; M.L. and J.E. performed research; A.C. and N.M. provided TOEFL data; M.L., A.C., and N.M. analyzed data; and M.L. and J.E. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2023 the Author(s). Published by PNAS. This article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

¹To whom correspondence may be addressed. Email: mollylewis@gmail.com or jevans@uchicago.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2300986120/-/DCSupplemental>.

Published December 11, 2023.

intuitive and common. By contrast, their linkage may seem strange and rare when domains are distant in semantic space. In this way, measured misalignment between close and distant concepts across language pairs would allow us to better characterize and predict the existence of “semantic accents” previously identified in bilingual speakers (31, 32), but on large scales across the world’s languages.

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, (e.g., 11, 25, 33) or a few (7, 34, 35) semantic domains. 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 conditions 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 (36, 37). Word embeddings provide a systems-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 are highly correlated with human judgments of semantic similarity and encapsulate and represent culture-specific biases with fidelity (38–44). We describe computed word embeddings as representing the semantic space of a language and explore the semantic distance between pairs of languages in this space by evaluating continuous distances between word pairs in both. We then operationalize semantic domains by clustering words based on their loadings on embedding dimensions to compare “local” (within cluster) versus “global” (across cluster) variability in semantics between languages. We note that our distinction between “local” and “global” refers only to semantic distances within languages, synonymous with meanings that cluster versus those that span the language, and has no relationship with geographical distance. A positive correlation in word distances between two languages suggests that the two languages manifest similar relationships between lexical meanings, which we take as evidence for semantic similarity between those languages.

Using word embeddings, we compare the structure of semantic space for 35 different languages that span 11 language families in two stages and with two complementary 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 *SI Appendix, Fig. S1*, 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 without assuming translation equivalents of word meanings, and 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

(*SI Appendix, Fig. S2*). The striking similarity of patterns between findings from these two datasets confirms that the semantics of one’s native language influences the semantics of one’s second language for bilingual speakers (31, 45–49), 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 the relationship between the semantic systems of different languages is principled. 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 *SI Appendix, Fig. S3*), 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 English grammatical or syntactic errors (see *SI Appendix, Fig. S4A*).

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. We hypothesized that the amount of variability for a particular semantic domain would vary across languages, but in a principled manner. Following Gentner et al. (50–53), we posit that semantic domains referring to meanings more perceptually available and concrete such as “food” and “body” will be more similar across languages, relative to domains more conceptual and abstract like “injustice” and “democracy”. This hypothesis has been 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 (54, 55).

To test the concreteness hypothesis, we estimated the concreteness of each word based on human judgments and partitioned them into 10 contiguous sets separated by rising concreteness thresholds (56) (*SI Appendix, Fig. S5*). These sets overlapped strongly and significantly with semantic clusters in both the TOEFL ($\chi^2(81) = 1538.1$; $P < 0.00001$) and Wikipedia word samples ($\chi^2(81) = 5144.1$, $P < 0.00001$, and Fig. 1A), far

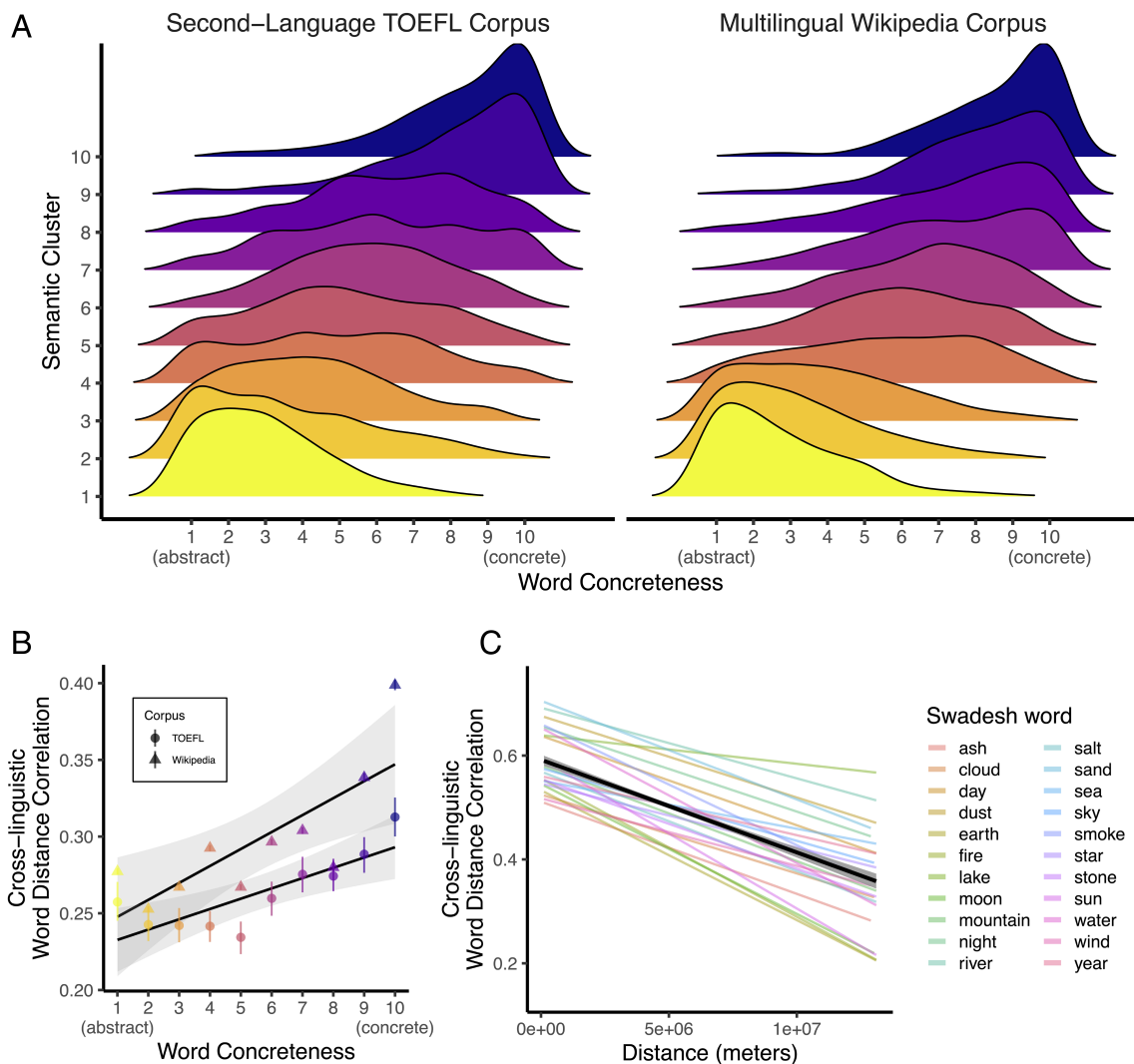


Fig. 1. (A) Distribution of words in each semantic cluster across concreteness deciles based on word embeddings obtained from Second-Language TOEFL Corpus and Multilingual Wikipedia Corpus. (B) Mean cross-linguistic word distance correlation (Pearson's r) as a function of the concreteness decile of the words. Larger values indicate more semantic similarity across languages. Point shape indicates corpus. Point ranges correspond to bootstrapped 95% CIs; range on model fit corresponds to the SE. (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. Colored lines correspond to the model fit for each word, and the black line shows the overall model fit and corresponding SE.

above what one would expect from a random distribution of concreteness over clusters. In line with the low semantic variability hypothesis, languages exhibit higher similarity in more perceptually concrete domains, and less in those more conceptually abstract (TOEFL: $r = 0.78$, $P = 0.008$; Wikipedia: $r = 0.82$; $P = 0.004$; Fig. 1B and see *SI Appendix, Figs. S6 and S7* for supporting analyses). Consistent with previous work, however, concreteness predicts only part of the cross-linguistic variability in semantic structure (12).

These results contrast with the findings of Thompson et al. (34) who find no relationship between the concreteness of word meanings and cross-linguistic semantic alignment using word embedding models. This distinction follows from their approach to sampling words and relationships, compared with our systems-level approach. We examined a large set of words (10,000) sampled randomly across the entire semantic space, rather than words hand-picked from a small list of semantic domains ($N = 21$). Further, our method for comparing word similarity cross-linguistically evaluates the relations between all words across semantic space, semantically near and far (5×10^7),

whereas their work only evaluated the 100 nearest-neighbors to target words (*SI Appendix, Fig. S8*). When we replicated their analysis, we find a small negative relationship between concreteness and semantic alignment, which grows larger and more strongly significant when our much larger collection of words and complete comparisons are considered.

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. (57) in their demonstration of a supposedly "universal" structure of lexical semantics (2016; e.g., "water," "sun," and "dirt") and still found moderate variability in pairwise-distances across languages.

This variability was highly predicted by physical and environmental distance. Languages in closer geographical proximity exhibit much more similar semantic representations for almost all

of these highly concrete words (QAP $P < 0.01$: physical: 20/22; Fig. 1C and *SI Appendix*, Fig. S9) and environmental disparity explains variations for some of the items (environmental: 6/22). This suggests that even for concrete meanings, there is substantial variability in the structure of semantics across languages and this variability can be predicted by a combination of differences in the perceptual experience of language speakers and their potential for direct or indirect cultural contact.

We then expand this analysis to explore the full semantic space. We predict this with a richer collection of differences including not only geographical and environmental similarity, but also lexical form and grammatical similarity, and cultural similarity comprising factors ranging from likeness in the structure of kinship, religion, politics, and social class. These analyses suggest that language semantics—even associated with concrete words—vary substantially across languages, but remain predictable by cultural difference and environmental distance (Fig. 2 and *SI Appendix*, Table S1 and Fig. S10).

Local Similarity and Global Variability in Semantic Space

These analyses and theory 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 A and B). 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 “sacred,” “deity,” etc.). Finally, meanings may differ evenly across the system (*SI Appendix*, Figs. S11–S13).

Following Rosch et al. (54), we anticipated that locally clustered meanings would remain largely conserved across languages. They articulate how “feathers,” “wings,” and “beaks”

frequently occur together” in the world, resulting in a common semantic cluster (54). Elsewhere Malt and colleagues demonstrate how “perception of stimulus properties by individuals interacts with linguistic and cultural histories, but their interaction is constrained by structure in the stimulus space” (7, 12). This perspective suggests that local clusters of meaning may be much more likely conserved across languages than farflung global associations that span the cultural system (e.g., 58).

Supporting this expectation, we first found that word pairs from the same concreteness decile (“local” relations) tended to cluster together and were more similar to each other across languages, relative to word pairs across concreteness deciles (“global” relations; TOEFL: $M = 0.025$, $SD = 0.005$; $t(594) = 128.64$; $P < 0.0001$; $d = 1.28$ [1.15, 1.4]; $W = 177310$, $P < 0.0001$; Wikipedia: $M = 0.035$, $SD = 0.015$; $t(594) = 56.88$; $P < 0.0001$; $d = 0.27$ [0.16, 0.39]; $W = 177272$, $P < 0.0001$; *SI Appendix*, Figs. S11–S13). Malt’s work, however, casts doubt that the distinction between concrete and abstract words would account for the greatest cross-language variation. Her work reveals substantial cross-language variation in labels of the seemingly concrete domain of “containers” (12, 24), just as others have found substantial cross-language variation in body parts (14, 35). This suggests that domain-clustered meanings may better account for variation in cross-linguistic semantic structure than the distinction between concrete and abstract concepts.

To directly measure local-global semantic relations, which are not exclusively organized by concreteness, we clustered words according to their position within semantic space, and examined the relationship between words within and between clusters across languages (see *Materials and Methods* for detail; *SI Appendix*, Fig. S14). 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 ($M = 0.058$, $SD = 0.008$; $t(594) = 185.97$; $P < 0.0001$; $d = 2.84$

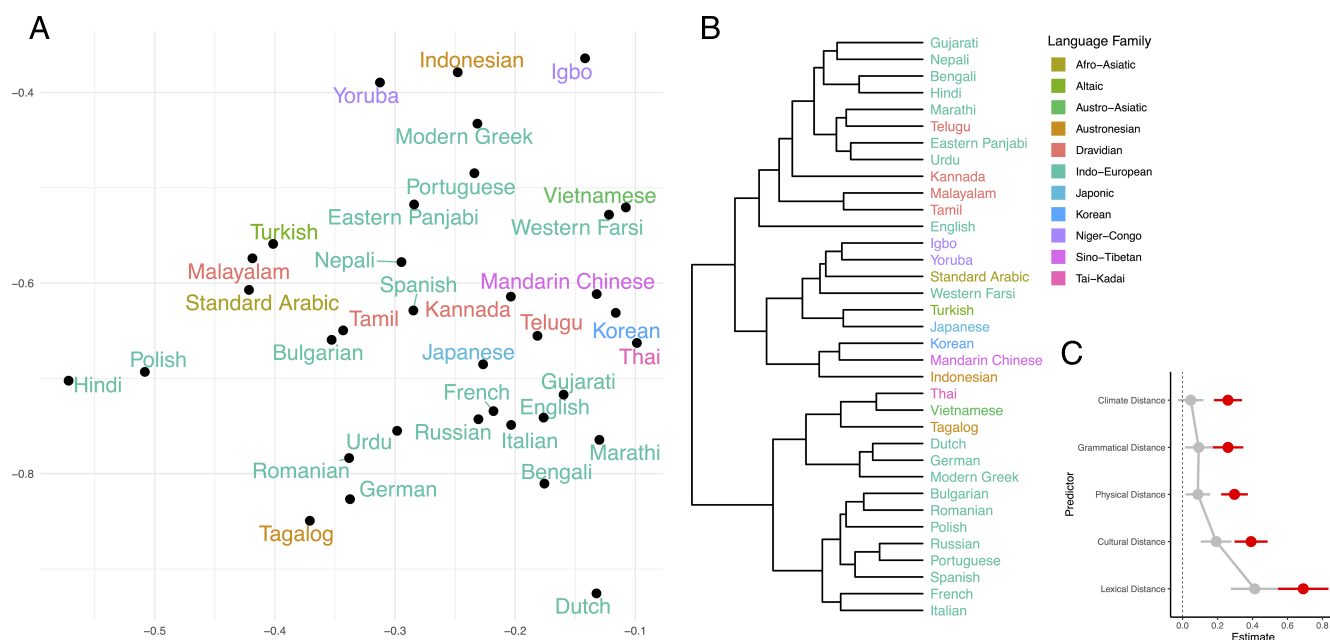


Fig. 2. (A) Two-dimensional projection of language centroids calculated from document embeddings of the Second Language TOEFL Corpus. Color corresponds to language family. (B) Hierarchical clustering of languages based on pairwise language distances of language centroids. (C) Standardized linear estimates of semantic distance predictors. Ranges are 95% CIs. Red points indicate estimates from the single-predictor model; gray points indicate estimates from the additive linear model with all five predictors included.

[2.68, 3]; $W = 177310$, $P < 0.0001$) and Wikipedia datasets ($M = 0.038$, $SD = 0.024$; $t(594) = 38.27$; $P < 0.0001$; $d = 0.31$ [0.19, 0.42]; $W = 174402$, $P < 0.0001$; Fig. 4A and *SI Appendix*, Figs. S15 and S16). This effect was not related to the grammatical similarity of the languages (*SI Appendix*, Figs. S4B and S17), showing remarkable consistency across languages (*SI Appendix*, Figs. S18–S20). Further, the effect grew substantially larger as the number of semantic clusters increased (Fig. 3D and *SI Appendix*, Fig. S21).

Implications of Cross-Linguistic Semantic Diversity

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. We interpret cross-linguistic variability in relations between semantic domains as creating measurable differences in the cognitive availability of linkages like associations, analogies, and metaphors that connect those domains. People who “think” in Greek cognitively follow and produce distinct semantic associations in language from others who “think” in Arabic, Farsi, Igbo, or Chinese. Our research also characterizes the existence and extent of “semantic accents” previously identified in bilingual speakers (31, 32) on large scales across the world’s languages.

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 (59), but because associations, metaphors, and stories interlink different domains of meanings in one language culture than another (58). 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.

Our findings invite future research to examine the cognitive consequences of cross-linguistic differences in semantic relations for non-native speakers. Previous research suggests how speakers of different languages may perceive (60, 61), remember (62–67), and learn (68, 69) about the world in different ways. Our work suggests that these effects may extend to tasks that require linking clusters of meanings. Future research could directly detect local-global semantic accents behaviorally using a classic priming paradigm (32), with native and non-native speakers producing different associates in proportion to the semantic alignment shown here. Our work further predicts differences between native and non-native speakers in language comprehension that transcend native familiarity (70–72). Global semantic relations deployed in analogies or metaphors should be more difficult to process in one’s second language relative to one’s native language if those parts of the language do not align. In contrast, language relying on local semantic relationships should be easier to process, regardless of language. This prediction could be tested using behavioral measures like reaction time or neural signals, like N400 (73), where we expect processing difficulty in proportion to the semantic distances we measure in this study (i.e., a larger N400 effect for more semantically distant cross-language word pairs).

Domains that lie close within a language as assessed by our models are associated with an increased likelihood that speakers with native fluency transition from one concept to another across the proximate domain boundary (74). Transitions occur through discursive pivots within conversations, expositions, and narratives. They also occur in comparisons, such as similes and metaphors. We leave to future work the task of documenting

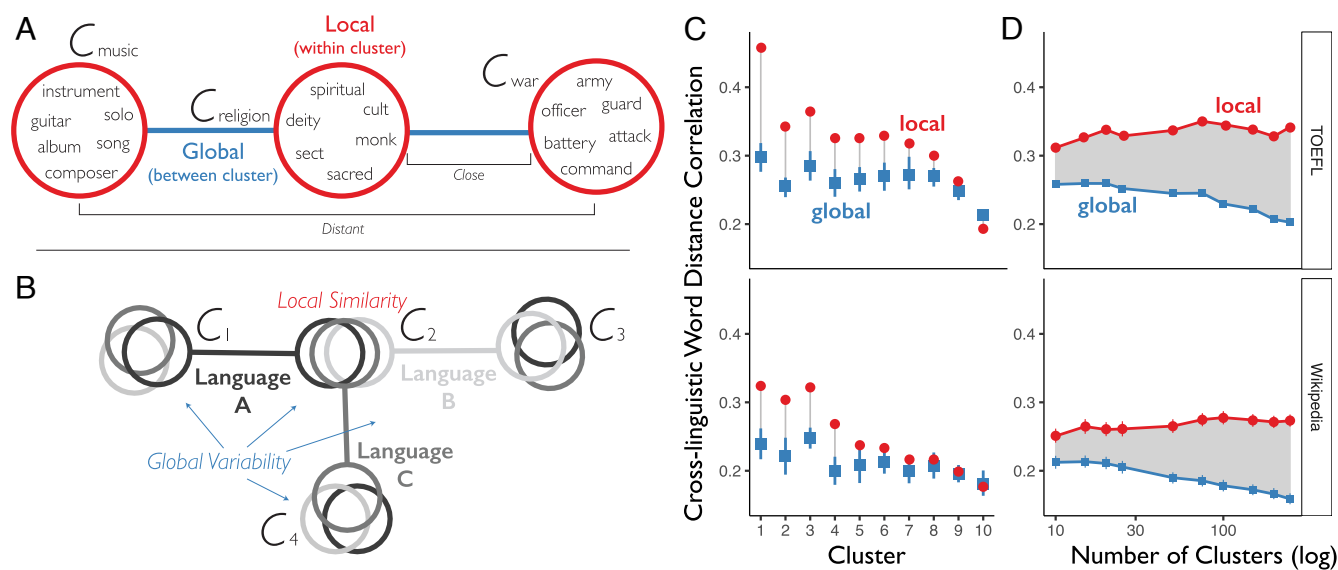


Fig. 3. (A) Schematic representation of local and global distances in word embedding models. The figure shows three “clusters” of meanings and the relationships between them. Local relationships are within-cluster distances; global relationships are between-cluster distances. Some clusters are closer 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% CIs.

whether specific types of transition (e.g., narrative turns vs. metaphors) become differentially likely with embedded proximity or whether all become more likely at a roughly equal rate as they are environmentally primed, reinforced, and become cognitively available to language speakers.

Our findings also have powerful practical implications for language learning—they suggest 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 suggest discrepancies between native and non-native speakers in both comprehension and production could be reduced by training on the semantic associations in one's second language. Moreover, they point to 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.

Understanding how languages semantically vary also provides justification for the importance of cultural difference. Our findings could allow teachers and translators to compensate for semantic surprise and smooth intercultural communication. Artists, scientists, and scholars could equally leverage them to elevate disconnection and produce surprise. Recent scholarship on research innovation demonstrates how expeditions of academic outsiders produce the most unexpected and impactful findings. Only outside their scholarly languages do “alien” researcher insights and tools have the potential to generate novel solutions to previously intractable problems (76). Collective cognitive diversity might emerge from collaboration among those with native languages containing divergent semantic structures (77). Not only creators, but anyone making complex decisions could benefit from targeted confusion, where thinking like a native is more liability than benefit.

One open question from our findings is the source of the cross-linguistic semantic differences we observe. For example, why do English and Persian differ in their semantic structure? In principle, two broad causal forces could lead to cross-linguistic variability. The first is variation in experience with the world, either from different physical environments or associations emergent from cultural life (78–81). For example, concepts of family and food might be more strongly associated in language cultures with resource abundance and the ongoing tradition of

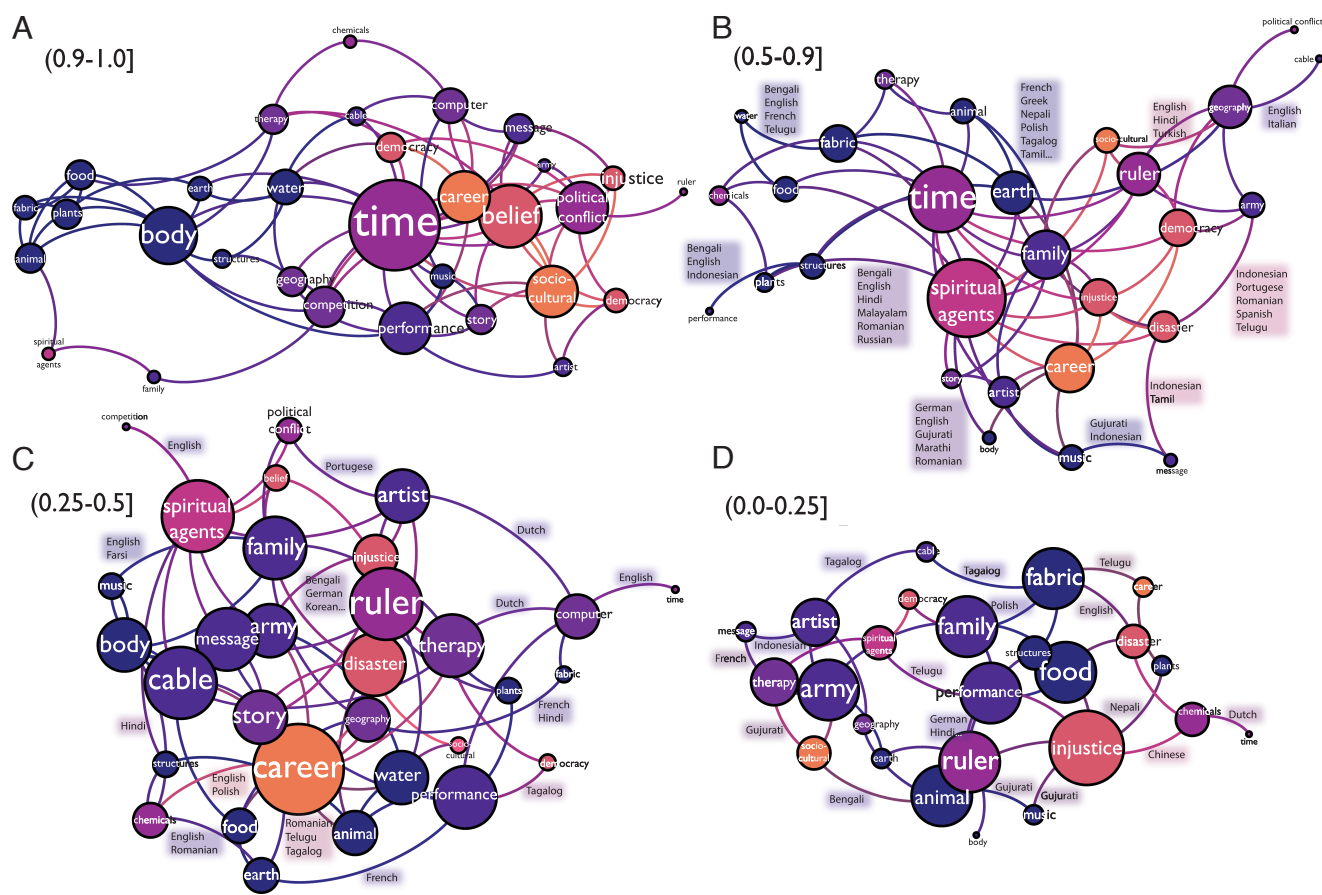


Fig. 4. Semantic cluster graphs where links between clusters represent whether or not those clusters are within the top quartile of global semantic closeness for the proportion of languages that share both clusters specified above each graph. Links between clusters in Panel A are shared by almost all (0.9-1.0) languages in which the clusters manifest. Panels B–D 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 a force-directed algorithm that draws together the most connected clusters. Clusters were labeled by the authors, and were represented in force-directed network layouts with Gephi (75).

family gathering to produce and consume food together. Other differences in experience follow from technological configurations and complexity, which also shape associations represented in language.

The second potential causal force is linguistic associations that violate experience. Consider homonymy, where two words with divergent meanings sound similar, or polysemy, where the same word holds different or divergent meanings. Homonymy and polysemy may arise through the merger of distinct language populations, or those previously separated where the dialects underwent functional evolution before re-merging. These historical “accidents” become apparent in the frequency of uncertain or competing etymologies, where word cognates and roots exist in multiple parent languages (e.g., Latin, Greek, and Saxon for English). Despite their “accidental” character, when similar or identical words link divergent meanings, they indirectly hyperlink other concepts and domains related to those linked (82). This path-dependent process may channel progressively stronger semantic associations in language, inspiring material proximities that reinforce the association. In this way, the two potential causal forces—experience with the world and linguistic accidents—are not mutually exclusive and may recursively influence one another over time. Once linguistic associations become available for language speakers, however, they likely become reinforced through linguistic experience (43, 83, 84).

Using experimentation to understand the ways in which causal forces interact to drive the semantic variability we observe is an important target for future work. Using an artificial language learning paradigm (85), one could create concepts that hyperlink multiple meanings (i.e., novel homonyms) or teach new knowledge that functionally links them (e.g., novel composite technologies or theories), then evaluate the degree to which each contributes to altering the perceived distance between meanings (68). One could further stage experiments that compare linguistic accidents with altered embodied experiences (55). Alternatively, analysts could take a historical observational approach and measure the influence of historical accidents such as language mergers compared with changes in the emergence of new knowledge and technology, giving rise to functional shifts in the structure of meaning (e.g., ref. 86).

Our analysis involves several limitations. First, we model language semantics in our analyses exclusively through word co-occurrence information captured through word embedding models. Prior work has demonstrated that estimates of meaning similarity derived from word embeddings are strongly correlated with explicit human judgments of word similarity (38, 44). Nevertheless, they do not perfectly correlate, suggesting that our method may not fully capture variability in cross-linguistic semantics. Contextual embeddings from the transformer models underlying modern chatbots (87, 88), provide greater resolution by capturing the semantics of words in context, but they require more language data than the modest corpora available to us. Second, our corpora—native language expository articles from Wikipedia and second language TOEFL essays—reveal that there is cross-linguistic variability in connotative semantic associations, but we do not have comparable samples of fiction, poetry, conversation, or informal text to directly test this hypothesis. Finally, our data are correlational: Behavioral experiments are needed to definitively establish the causal role of semantic space on the cognitive accessibility of semantic associations.

Understanding how languages differ in their semantics has the potential to facilitate better cross-cultural communication, and provide justification for the importance of cultural difference.

Our work describes the general principles governing this variability and characterizes these differences across semantic space.

Materials and 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 essay prompts. The essays were written by equal number of participants from 35 different languages. Each essay was associated with a 1 to 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 (89) 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 (90). 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 *SI Appendix, Fig. S3*) 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 (36). 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.

Word-Level Analyses. The conceptual concreteness of a word was estimated using previously collected human judgments (56). 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 (91) to cluster the words into 10 clusters each based on their semantics, through 50 iterations. To compare groupings by concreteness deciles and semantic clusters, we report in the main text the χ^2 statistic of word counts in a *N*-cluster by concreteness decile matrix (10×10 ; Fig. 1A). We also performed clustering for many other solutions, from 10 to 250 clusters, as shown in Fig. 3D, with greater differences between local and global correlations

as cluster number grew. In Fig. 3D, clusters were determined based on the model trained on English Wikipedia, using word loadings on the 300 Wikipedia embedding dimensions for all pairs of languages. We also created clusters based on the Wikipedia entries for every language in our sample ($N = 35$).

We next evaluated the semantic similarity of words across languages as a function of word concreteness (Fig. 1B). We calculated the pairwise distance (cosine; see *Materials and Methods* below) 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 used the same set of words as above to compare the pairwise cosine distances between words in different concreteness deciles ("global") to those in the same concreteness decile ("local," described above). To measure global alignment, we calculated 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.). To measure local alignment, we calculated the pairwise distance between words in the same concreteness deciles, and then calculated the correlation for each language pairing and decile (1-1, 2-2, 3-3, etc.). We then compared the mean local correlation to the mean global correlation across language pairs ($N = 595$). We report statistics for both parametric (paired t -test) and non-parametric (paired Wilcoxon signed-rank test) analyses.

We conducted a parallel analysis using word sets defined by the semantic clusters (described above) rather than concreteness deciles, varying the number of clusters considered (10 to 250). In this analysis, "local" refers to within-cluster distances and "global" refers to across-cluster distances. 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% CI.

Finally, we replicated the cluster-based analysis using clusters determined by native language embeddings, rather than English-based clusters (*SI Appendix, Fig. S21*). We performed pairwise comparisons of word distances within and across cluster boundaries, calculating this for TOEFL essays in both languages based on the cluster solution from each of the two languages, then averaging. 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. This manifests the same pattern of results as those derived from English clusters suggesting the insensitivity of our findings to distinct sources of semantic cluster structure. This, of course, reinforces our finding that local structure is much more similar across languages than global structure.

Semantic Similarity in Swadesh Words. We used the Google Translate API to translate the 22 words analyzed by Youn et al. (92) (a subset of the Swadesh list) into each of our target 35 languages. We included the variants analyzed 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 (36). 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 (93) 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) (94). The QAP

procedure estimates P -values in a way that accounts for the non-independence of observations (see *Materials and Methods* below).

Climate similarity was based on climate data obtained from WorldClim (95) on the basis of geographical coordinates from Glottolog 2.7 (93). 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 ref. 96. Grammatical distance between languages is based on similarity of 130 typological features for each language coded from the WALS database (97). Lexical similarity is based on the Levenshtein edit distance between the phonological forms of a standard set of 40 words in each language (98) (ASJP16). Finally, cultural similarity is based on data from D-place, an ethnographic atlas of cultural traits (34, 99). 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 *SI Appendix, Fig. S10B* 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 (42). 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 hypersphere's dimension approaches infinity, its volume approaches zero.

A geometric interpretation may be preferable to a probability one like the Kullback-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 (100).

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 (101). 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 α 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.

Data, Materials, and Software Availability. Data and code are available through the GitHub repository associated with the project: <https://github.com/millewis/SYSTEMSEM> (102). Some study data available Personal TOEFL Essays can only be analyzed in a secure setting because they contain personally identifying information.

ACKNOWLEDGMENTS. We acknowledge NSF #1520074 for partial support for this project, and thank the Educational Testing Service (A.C. and N.M.) for sharing and processing TOEFL data.

Author affiliations: ^aPsychology & Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213; ^bDataminr, Inc., New York, NY 10016; ^cEducational Testing Service, Princeton, NJ 08541; ^dSociology & Data Science, University of Chicago, Chicago, IL 60637; and ^eSanta Fe Institute, Santa Fe, NM 87501

1. P. Kay, T. Regier, Language, thought and color: Recent developments. *Trends Cognit. Sci.* **10**, 51–54 (2006).
2. N. Zaslavsky, C. Kemp, N. Tishby, T. Regier, Color naming reflects both perceptual structure and communicative need. *Top. Cognit. Sci.* **11**, 207–219 (2019).
3. T. Regier, P. Kay, Language, thought, and color: Whorf was half right. *Trends Cognit. Sci.* **13**, 439–446 (2009).
4. E. Wnuk, A. Verkerk, S. C. Levinson, A. Majid, Color technology is not necessary for rich and efficient color language. *Cognition* **229**, 105223 (2022).
5. A. M. Majid, M. v. S. Bowerman, J. S. Boster, The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cogn. Linguist.* **18**, 133–152 (2007).
6. J. C. Jackson *et al.*, Emotion semantics show both cultural variation and universal structure. *Science* **366**, 1517–1522 (2019).
7. Y. Xu, K. Duong, B. C. Malt, S. Jiang, M. Srinivasan, Conceptual relations predict colexification across languages. *Cognition* **201**, 104280 (2020).
8. D. Gentner, M. Bowerman, Why some spatial semantic categories are harder to learn than others: The typological prevalence hypothesis. *Crossling. Appr. Psychol. Lang.: Res. Trad. Dan Isaac Slobin* **4652009**, 480 (2009).
9. C. Kemp, T. Regier, Kinship categories across languages reflect general communicative principles. *Science* **336**, 1049–1054 (2012).
10. A. Majid, M. Van Staden, Can nomenclature for the body be explained by embodiment theories? *Top. Cognit. Sci.* **7**, 570–594 (2015).
11. T. Regier, P. Kay, N. Khetarpal, Color naming reflects optimal partitions of color space. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1436–1441 (2007).
12. B. C. Malt, S. A. Sloman, S. P. Gennari, Universality and language specificity in object naming. *J. Mem. Lang.* **49**, 20–42 (2003).
13. M. Bowerman, S. Choi, "Space under construction: Language-specific spatial categorization in first language acquisition." in *Language in Mind: Advances in the Study of Language and Thought*, D. Gentner, S. Goldin-Meadow, Eds. (MIT Press, Cambridge, 2003), pp. 387–427.
14. S. C. Levinson, Parts of the body in Yéli Dnye, the Papuan language of Rossel island. *Lang. Sci.* **28**, 221–240 (2006).
15. B. C. Malt, S. A. Sloman, S. P. Gennari, "Speaking versus thinking about objects and actions." in *Language in Mind: Advances in the Study of Language and Thought*, D. Gentner, S. Goldin-Meadow, Eds. (MIT Press, Cambridge, 2003), pp. 81–112.
16. T. Regier, Whorf was half right. PsyCEXTRA Dataset. APA PsycNet. <https://doi.org/10.1037/e527312012-204>. Accessed 15 December 2021.
17. S. P. Gennari, S. A. Sloman, B. C. Malt, W. T. Fitch, Motion events in language and cognition. *Cognition* **83**, 49–79 (2002).
18. S. Choi, L. McDonough, M. Bowerman, J. M. Mandler, Early sensitivity to language-specific spatial categories in English and Korean. *Cognit. Dev.* **14**, 241–268 (1999).
19. A. Majid *et al.*, Differential coding of perception in the world's languages. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 11369–11376 (2018).
20. A. Majid, N. Burenhult, Odors are expressible in language, as long as you speak the right language. *Cognition* **130**, 266–270 (2014).
21. J. L. Huismann, R. van Hout, A. Majid, Patterns of semantic variation differ across body parts: Evidence from the Japonic languages. *Cognit. Linguist.* **32**, 455–486 (2021).
22. T. Regier, A. Carstensen, C. Kemp, Languages support efficient communication about the environment: Words for snow revisited. *PLoS One* **11**, e0151138 (2016).
23. D. Kemmerer, *Concepts in the Brain: The View from Cross-Linguistic Diversity* (Oxford University Press, 2019).
24. B. C. Malt, S. A. Sloman, S. Gennari, M. Shi, Y. Wang, Knowing versus naming: Similarity and the linguistic categorization of artifacts. *J. Memory Lang.* **40**, 230–262 (1999).
25. B. Berlin, P. Kay, *Basic Color Terms: Their Universality and Evolution* (University of California Press, 1991).
26. G. P. Murdock, Kin term patterns and their distribution. *Ethnology* **9**, 165 (1970).
27. F. Saussure, *Course in General Linguistics* (Peter Owen, London, 1916, 1960).
28. C. Lévi-Strauss, *Structural Anthropology* (Basic Books, 2008).
29. L. Marti, S. Wu, S. T. Piantadosi, C. Kidd, Latent diversity in human concepts. *Open Mind* **7**, 79–92 (2023).
30. G. Lupyan, "Whorf for the 21st century: From interactive processing to linguistic relativity." in *Proceedings of the Annual Meeting of the Cognitive Science Society*, M. Knauff, M. Pauen, N. Sebanz, I. Wachsmuth, Eds. (cognitivesciencesociety.org, 2013), 1678–1679.
31. E. Matsuki, Y. Hino, D. Jared, Understanding semantic accents in Japanese-English bilinguals: A feature-based approach. *Bilingual.: Lang. Cognit.* **24**, 137–153 (2021).
32. Y. Dong, S. Gui, B. MacWhinney, Shared and separate meanings in the bilingual mental lexicon. *Bilingual.: Lang. Cognit.* **8**, 221–238 (2005).
33. R. Baddeley, D. Attewell, The relationship between language and the environment information theory shows why we have only three lightness terms. *Psychol. Sci.* **20**, 1100–1107 (2009).
34. B. Thompson, S. G. Roberts, G. Lupyan, Cultural influences on word meanings revealed through large-scale semantic alignment. *Nat. Hum. Behav.* **4**, 1029–1038 (2020).
35. A. Majid, F. Jordan, M. Dun, Semantic systems in closely related languages. *Lang. Sci.* **49**, 1–18 (2015).
36. P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information. *TACL* **5**, 135–146 (2017).
37. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. arXiv [Preprint] (2013). <https://arxiv.org/pdf/1301.3781.pdf> (Accessed 10 January 2014).
38. F. Hill, R. Reichart, A. Korhonen, Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Comput. Linguist.* **41**, 665–695 (2015).
39. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
40. N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E3635–E3644 (2018).
41. T. Bolukbasi, K. W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Adv. Neur. Inf. Process. Syst.* **29**, 4349–4357 (2016).
42. A. C. Kozlowski, M. Taddy, J. A. Evans, The geometry of culture: Analyzing the meanings of class through word embeddings. *Am. Sociol. Rev.* **84**, 905–949 (2019).
43. M. Lewis, G. Lupyan, Gender stereotypes are reflected in the distributional structure of languages. *Nat. Hum. Behav.* **4**, 1021–1028 (2020).
44. F. Pereira, S. Gershman, S. Ritter, M. Botvinick, A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognit. Neuropsychol.* **33**, 175–190 (2016).
45. E. Ameel, B. C. Malt, G. Storms, F. Van Assche, Semantic convergence in the bilingual lexicon. *J. Memory Lang.* **60**, 270–290 (2009).
46. B. C. Malt, P. Li, A. Pavlenko, H. Zhu, E. Ameel, Bidirectional lexical interaction in late immersed Mandarin-English bilinguals. *J. Memory Lang.* **82**, 86–104 (2015).
47. E. Ameel, G. Storms, B. C. Malt, S. A. Sloman, How bilinguals solve the naming problem. *J. Memory Lang.* **53**, 60–80 (2005).
48. S. Jarvis, A. Pavlenko, *Crosslinguistic Influence in Language and Cognition* (Routledge, 2008).
49. A. Pavlenko, S. Jarvis, Bidirectional transfer. *Appl. Linguist.* **23**, 190–214 (2002).
50. D. Gentner, Some interesting differences between verbs and nouns. *Cognit. Brain Theory* **4**, 161–178 (1981).
51. D. Gentner, "Why nouns are learned before verbs: Linguistic relativity versus natural partitioning" in *Language Development: Syntax and semantics*, S. A. Kuczaj, Ed. (Psychology Press, 1982).
52. D. Gentner, "Why Verbs Are Hard to Learn." in *Action Meets Word: How Children Learn Verbs*, K. Hirsh-Pasek, R. M. Golinkoff, Eds. (Oxford University Press, 2006), pp. 544–564.
53. Y. Zhou, D. Yurovsky, A common framework for quantifying the learnability of nouns and verbs. *Proc. Annu. Meet. Cogn. Sci. Soc.* **43**, 314–320 (2021).
54. E. Rosch, C. B. Mervis, Family resemblances: Studies in the internal structure of categories. *Cognit. Psychol.* **7**, 573–605 (1975).
55. D. Casasanto, Different bodies, different minds: The body specificity of language and thought. *Curr. Direct. Psychol. Sci.* **20**, 378–383 (2011).
56. M. Brysbaert, A. B. Warriner, V. Kuperman, Concrete ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **46**, 904–911 (2014).
57. H. Youn *et al.*, On the universal structure of human lexical semantics. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 1766–1771 (2016).
58. G. Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind* (University of Chicago Press, 2008), pp. 58–67.
59. W. V. O. Quine, *Word and Object* (MIT Press, 2013).
60. J. Winawer *et al.*, Russian blues reveal effects of language on color discrimination. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7780–7785 (2007).
61. B. Boutonnet, G. Lupyan, Words jump-start vision: A label advantage in object recognition. *J. Neurosci.* **35**, 9329–9335 (2015).
62. E. F. Loftus, J. C. Palmer, Reconstruction of automobile destruction: An example of the interaction between language and memory. *J. Verb. Learn. Verb. Behav.* **13**, 585–589 (1974).
63. A. S. LaTourrette, S. R. Waxman, Naming guides how 12-month-old infants encode and remember objects. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 21230–21234 (2020).
64. M. C. Frank, D. L. Everett, E. Fedorenko, E. Gibson, Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition* **108**, 819–824 (2008).
65. A. Majid, M. Bowerman, S. Kita, D. B. Haun, S. C. Levinson, Can language restructure cognition? The case for space. *Trends Cognit. Sci.* **8**, 108–114 (2004).
66. S. C. Levinson, *Space in Language and Cognition: Explorations in Cognitive Diversity* (Cambridge University Press, 2003), vol. 5.
67. A. Pavlenko, "Thinking and speaking in two languages: Overview of the field" in *Thinking and Speaking in Two Languages* (2011), pp. 237–257.
68. G. Lupyan, D. H. Rakison, J. L. McClelland, Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychol. Sci.* **18**, 1077–1083 (2007).
69. M. Zettersten, G. Lupyan, Finding categories through words: More nameable features improve category learning. *Cognition* **196**, 104135 (2020).
70. S. Hayakawa, B. Keysar, Using a foreign language reduces mental imagery. *Cognition* **173**, 8–15 (2018).
71. A. Costa *et al.*, Your morals depend on language. *PLoS One* **9**, e94842 (2014).
72. B. Keysar, S. L. Hayakawa, S. G. An, The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychol. Sci.* **23**, 661–668 (2012).
73. M. Kutas, K. D. Federmeier, Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annu. Rev. Psychol.* **62**, 621–647 (2011).
74. S. Arora, Y. Li, Y. Liang, T. Ma, A. Risteski, Random walks on context spaces: Towards an explanation of the mysteries of semantic word embeddings. arXiv [Preprint] (2015). <http://arxiv.org/abs/1502.03520> (Accessed 11 March 2020).
75. M. Bastian, S. Heymann, M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks." in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, W. Cohen, N. Nicolov, Eds. (AAAI, 2009), pp. 361–362.
76. F. Shi, J. Evans, Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nat. Commun.* **14**, 1641 (2023).
77. R. B. Freeman, W. Huang, Collaboration: Strength in diversity. *Nature* **513**, 305–305 (2014).
78. R. Dale, G. Lupyan, Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Adv. Complex Syst.* **15**, 1150017 (2012).
79. G. Lupyan, R. Dale, "The role of adaptation in understanding linguistic diversity" in *The Shaping Language: The Relationship between Structures Languages their Social, Cultural, Historical, Natural Environments* (2015), pp. 289–316.
80. G. Lupyan, R. Dale, Why are there different languages? The role of adaptation in linguistic diversity. *Trends Cognit. Sci.* **20**, 649–660 (2016).
81. G. Lupyan, R. Dale, Language structure is partly determined by social structure. *PLoS One* **5**, e8559 (2010).
82. P. Aceves, J. A. Evans, Human languages with greater information density increase communication speed, but decrease conversation breadth. arXiv [Preprint] (2021). <https://arxiv.org/pdf/2112.08491.pdf>. Accessed 15 December 2021.
83. A. G. Greenwald, An AI stereotype catcher. *Science* **356**, 133–134 (2017).

84. M. L. Lewis, M. C. Frank, The length of words reflects their conceptual complexity. *Cognition* **153**, 182–195 (2016).
85. S. Kirby, H. Cornish, K. Smith, Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 10681–10686 (2008).
86. M. Pagel, Q. D. Atkinson, A. Meade, Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–720 (2007).
87. A. Vaswani *et al.*, Attention is all you need. *Adv. Neur. Inf. Process. Syst.* **30**, 1–11 (2017).
88. A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving language understanding by generative pre-training (2018). <https://www.mikecaptain.com/resources/pdf/GPT-1.pdf>. Accessed 5 January 2019.
89. Q. Le, T. Mikolov, "Distributed representations of sentences and documents" in *International Conference on Machine Learning* (2014), pp. 1188–1196.
90. R. Rehurek, P. Sojka, Gensim-Python framework for vector space modelling. *NLP Centre Faculty Inf. Masaryk Univ. Brno, Czech Repub.* **3** (2011).
91. J. A. Hartigan, M. A. Wong, Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **28**, 100–108 (1979).
92. 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).
93. H. Hammarström, S. Nordhoff, Langdoc: Bibliographic infrastructure for linguistic typology. *Oslo Stud. Lang.* **3**, 31–43 (2011).
94. C. T. Butts, SNA: Tools for Social Network Analysis (2016). R package version 2.4.
95. S. E. Fick, R. J. Hijmans, Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **37**, 4302–4315 (2017).
96. D. Dediu, Making genealogical language classifications available for phylogenetic analysis: Newick trees, unified identifiers, and branch length. *Lang. Dyn. Change* **8**, 1–21 (2018).
97. M. S. Dryer, M. Haspelmath, Eds., *WALS Online* (Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013). <http://wals.info>.
98. S. Wichmann *et al.*, The ASJP database (version 16) (2013). <http://asjp.cldd.org/>.
99. K. R. Kirby *et al.*, D-place: A global database of cultural, linguistic and environmental diversity. *PLoS One* **11**, e0158391 (2016).
100. O. Levy, Y. Goldberg, "Neural word embedding as implicit matrix factorization" in *Advances in Neural Information Processing Systems* (2014), pp. 2177–2185.
101. D. Dekker, D. Krackhardt, T. A. Snijders, Sensitivity of MROAP tests to collinearity and autocorrelation conditions. *Psychometrika* **72**, 563–581 (2007).
102. M. Lewis, Code and processed data for main analyses of "Local similarity and global variability characterize the semantic space of human languages". Local-Global Semantics PNAS. <https://github.com/mllewis/SYSTEMSEM>. Deposited 3 May 2022.