When do languages use the same word for different meanings? The Goldilocks Principle in colexification

Thomas Brochhagen[a,1] **and Gemma Boleda**[a,b]

[a]Department of Translation and Language Sciences, Universitat Pompeu Fabra, Roc Boronat, 138, 08018, Barcelona, Spain; [b]Catalan Institution for Research and Advanced Studies (ICREA), Passeig Lluís Companys, 23, 08010, Barcelona, Spain

This manuscript was compiled on August 20, 2021

It is common for languages to express multiple meanings with the same word, a phenomenon known as colexification. For instance, the meanings FINGER and TOE colexify in the word *dedo* in Spanish, while they do not colexify in English. Colexification has been suggested to follow universal constraints. In particular, previous work has shown that related meanings are more prone to colexify. This tendency has been explained in terms of the cognitive pressure for simplicity, since expressing related meanings with the same word makes lexicons easier to learn and use. The present study examines the interplay between this pressure and a competing universal constraint, the functional pressure for languages to maximize informativeness. We hypothesize that colexification follows a Goldilocks principle: meanings are more likely to colexify if they are related (fostering simplicity), but not so related as to become confusable and cause misunderstandings (fostering informativeness). We find support for this principle in data from over 1200 languages and 1400 meanings. Our results thus suggest that universal principles shape the lexicons of natural languages. More broadly, they contribute to the growing body of evidence suggesting that languages evolve to strike a balance between competing functional and cognitive pressures.

language universals | colexification | cognitive effort | ambiguity | efficient communication

The association of multiple meanings with the same form is pervasive across natural languages (1–4). For instance, as illustrated in Figure 1A, the Spanish word *dedo* can refer to both a finger and a toe. That is, Spanish **colexifies** these two meanings, using a single word to express both (5). English instead expresses them with different words, *finger* and *toe*. Some colexifications are rather idiosyncratic; having a word referring to both a fruit and an alkaline substance, such as English *lime*, is not common. Others, instead, are attested throughout the world (5–9). The colexification of TOE and FINGER, for example, is found in at least 135 languages (10). These languages are spoken in different parts of the globe, and many are phylogenetically unrelated. This suggests that universal forces are at play, giving rise to systematic cross-linguistic patterns.

This study investigates how the interplay between two major forces shapes the lexical structure of natural languages, using large-scale cross-linguistic colexification data (10). The first force is the cognitive pressure for **simplicity**. A number of studies suggest that aspects of languages that are easier to learn and use will tend to be favored over time (11–13). In the extreme, the simplest language would colexify all meanings, using a single word form to express them all. However, while simple to learn, this would not be very useful from a communicative point of view. Indeed, a competing force drives languages to complexity: the need for them to be **informative**, in the sense of supporting accurate information transfer (14–21). In the other extreme, then, a maximally informative lexicon would have one word per meaning. But this would create immense lexicons that would be difficult to learn and use.

A growing body of research argues that languages are efficient in the sense that they strike a good balance between informativeness and simplicity (19–24). Natural language lexicons, in particular, have been suggested to be shaped by a trade-off between the two (20); however, so far only restricted domains have been explored, such as color (25) and kinship (26). We here examine the potential interaction between simplicity and informativeness in the lexicon at a larger scale, covering over 1400 meanings and more than 1200 languages.

We build on recent work that suggests that related meanings, like FINGER and TOE, tend to be expressed by the same word more than unrelated meanings, like CITRUS FRUIT and ALKALINE SUBSTANCE (9, 27). This tendency has been attributed to pressure for simplicity. The structure of lexicons as well as semantic memory may favor the colexification of meanings that are easy to relate to one another. This has been argued to assist vocabulary acquisition (with established word-meaning associations providing a scaffold for new meanings), as well as lexical retrieval and interpretation (6, 9, 28).

Significance Statement

While the way meanings are expressed varies substantially across languages, some are expressed by the same word more often than others. We find evidence that this is due to language evolving to strike a balance between two competing forces: simplicity and informativeness. More related meanings are more likely to be assigned to the same word; fostering simplicity. However, this tendency is counteracted by a communicative need to distinguish meanings in context. If meanings are so similar as to be confusable, they will be less likely to be conflated; fostering informativeness. These results suggest that, while different linguistic communities may care to distinguish different sets of meanings, the interaction of simplicity and informativeness universally shapes the lexicons of natural languages.

T.B. and G.B. designed research; analyzed results; and wrote the paper. T.B. performed research.

The authors declare no competing interests.

[1]To whom correspondence should be addressed. E-mail: thomas.brochhagen@gmail.com

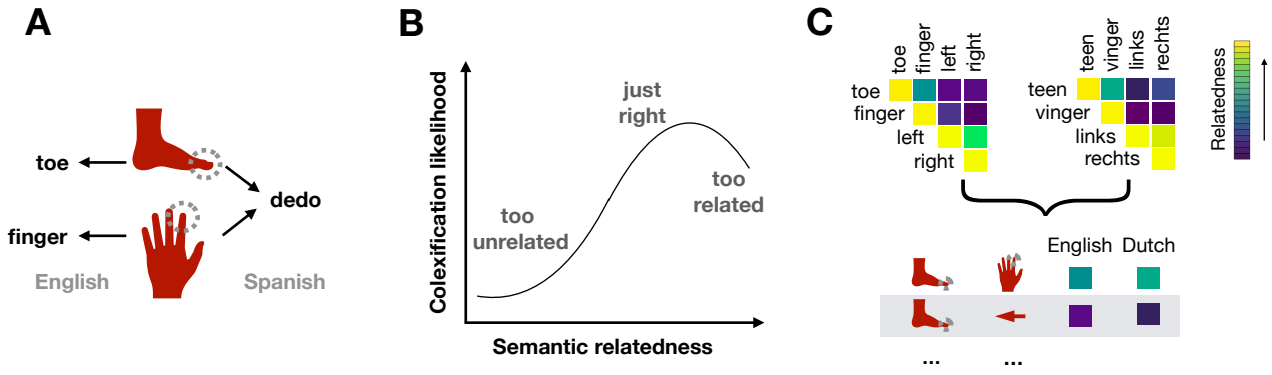

Fig. 1. A: Illustration of cross-linguistic differences in colexification. The meanings TOE and FINGER are expressed by the same word in Spanish (*dedo*) but not in English. B: Hypothesized relationship between the relatedness of meanings and the likelihood that they are expressed by the same form. The extremes on the x-axis are areas where meanings are either too related or too unrelated to be colexified. Weakly related meanings (e.g., ALKALINE SUBSTANCE and CITRUS FRUIT) are expected to be less likely to be expressed by the same form because they are hard to associate. Strongly related meanings (e.g., LEFT and RIGHT) are expected to be less likely to colexify because they are hard to tease apart in context. The middle-to-high range is conversely hypothesized to be particularly conducive to colexification. Meanings in this range (e.g., TOE and FINGER) may be easier to associate while not being too confusable in context. C: Estimation of relatedness between meanings. English and Dutch words are used as surrogates for meanings. Values for measures of semantic relatedness, such as distributional similarity, are computed on word pairs and used for meaning pairs. For instance, the distributional similarity of words *teen* and *vinger* in Dutch (upper right part of the figure) serves as a proxy for the similarity of meanings TOE and FINGER (lower part of the figure). The distributional similarity of the corresponding English words is taken as another estimate of the relatedness of these meanings.

We hypothesize that informativeness sets a limit to the tendency to colexify related meanings (cf. (27)): If meanings are too related, then expressing them with the same form can be disadvantageous from a communicative point of view. For instance, LEFT and RIGHT are highly related but express opposites. Using the same form for both can consequently lead to communicative failure (e.g. when giving directions). By contrast, CITRUS FRUIT and ALKALINE SUBSTANCE are unlikely to be mistaken by one another when uttering *lime* in any given context. Since the possibility to contextually disambiguate meanings is crucial for the persistence of lexical ambiguity (29–31), we expect there to be a point at which meanings are too related to be expressed by the same form. More specifically, we hypothesize that colexification follows a **Goldilocks principle**: meanings colexify if they are neither too unrelated, nor too related, but, as in the fairy tale *Goldilocks and the Three Bears*, "just right". This principle is illustrated in Figure 1B. Crucially, following the hypothesis that what hinders communication is meaning confusability in context (29, 30), we expect "too related" to mean "too confusable". In other words, we expect colexification likelihood to decrease in cases of high relatedness where confusability is at stake.

We find support for the Goldilocks principle in two analyses. The first uses data-induced measures of semantic relatedness to predict how likely meanings are to colexify. As hypothesized, we find that colexification likelihood increases with semantic relatedness, until an inflection point is reached for highly related meanings, after which the likelihood decreases. The second analysis further probes the role of confusability in decreasing colexification likelihood. We find that meanings that appear in very similar contexts, and particularly those that express opposites, are indeed less prone to colexify than other kinds of related meanings. Taken together, these findings thus support the hypothesis that natural language lexicons evolve to strike a balance between competing pressures for simplicity and informativeness.

Semantic relatedness follows the Goldilocks principle

To study the relationship between semantic relatedness and colexification, we fit a number of generalized additive logistic regressions (32) to colexification data spanning over 1200 languages and more than 1400 meanings, totaling 203,056 data points. This data comes from CLICS^3 (10), the largest cross-linguistic database of colexifications available to date. Data processing steps are detailed in *Materials and Methods* and SI Section 1.

Models. Our models characterize how likely a given pair of meanings is to colexify in a given language (e.g., TOE and FINGER in Spanish) as a function of one of three data-induced estimates of semantic relatedness, specified below. Comparing different kinds of estimates allows us to not only ask whether colexification follows the Goldilocks curve (Figure 1B), but also which kind of semantic information best characterizes the data.

Since language contact –facilitated by geographic proximity– and common linguistic ancestry influence colexification (8, 9), the models are also passed information about how often a pair of meanings colexifies in other languages. This information is weighted by the phylogenetic or geographic distance to the response language. Model details are given in *Materials and Methods* and SI Section 3.

Estimating semantic relatedness. We follow previous work in using words as surrogates for meanings when estimating semantic relatedness (e.g., (9, 27, 33)). More specifically, we use words in Dutch and English because large-scale data for the kinds of semantic resources that we focus on is available for them. As illustrated in Figure 1C, the relatedness of word pairs, such as *teen-vinger* in Dutch or *toe–finger* in English, are used as an estimate for the relatedness of their meanings (TOE-FINGER). These estimates are then used to predict the colexification likelihood of two meanings in other languages, factoring in geographic and phylogenetic influences.

We evaluate three measures of semantic relatedness: distributional similarity, associativity, and the first principal

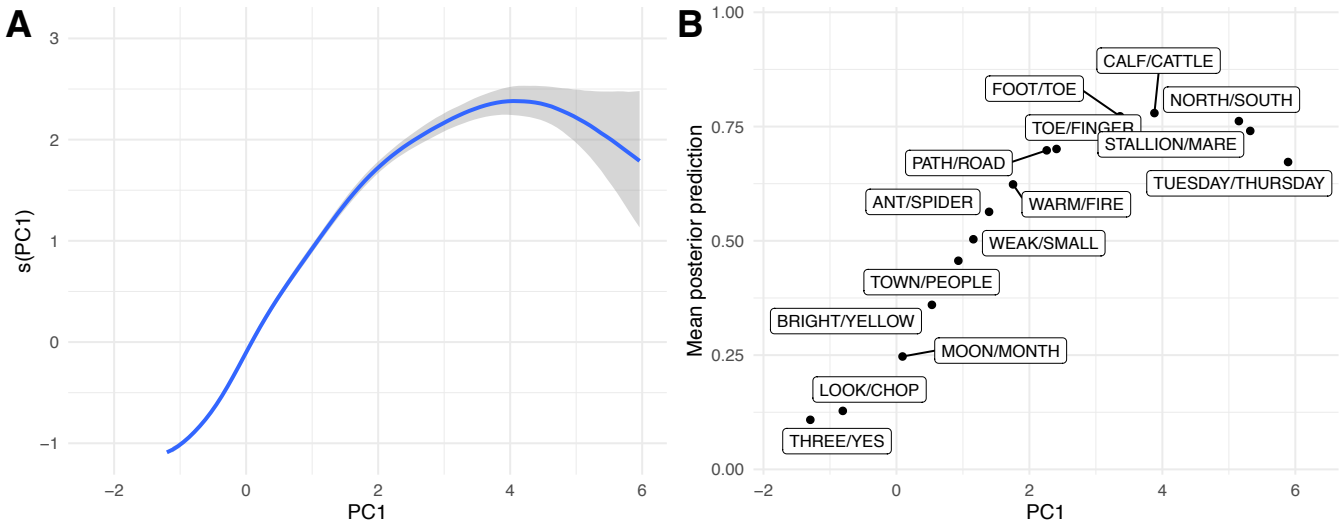

Fig. 2. A: Marginal effects of the best measure of semantic relatedness (PC1, in standardized units). Shading shows 95% credible intervals. The smooth function $s(\cdot)$ is inferred from the data and characterizes how PC1's contribution to colexification likelihood changes across values (on the logit scale). Uncertainty increases with deviation from the predictor's mean. This is expected given that data in this region is comparatively sparse. B: Example of mean posterior predictions for meaning pairs across standardized PC1 values estimated from Dutch words. Phylogenetic and geographic indicators were set to the minimum values they take in the data. These predictions are consequently about meaning pairs in a hypothetical language that has no nearby languages colexifying them.

component of these two measures. Each captures different facets of meaning. How well they characterize the data will thus also inform us about the kind of information that matters most for colexification.

Distributional similarity measures how similar the contexts of use of different linguistic expressions are, quantifying their contextual overlap based on large amounts of data, typically text corpora (34–37). To illustrate, the contexts of use of *left* and *right* are quite similar (distributional similarity of 0.57 in the English model that we use, with 1 being the maximum); *toe* and *finger* are also quite similar but less so (0.47); and *toe* and *left* are, expectedly, the least similar of these pairs (0.24). Details on distributional models are given in SI Section 1.B.

Associativity is derived from large-scale association norms (38, 39), obtained by asking subjects to produce words in response to a cue. For instance, when prompted by the word *toe*, a given subject may produce *foot*, *finger*, or *nail*. The *Material and Methods* and SI Section 1.C detail how measures of relatedness are derived from association norms. Using the examples from above and the English norms that we use in this study, the associativity of *left* and *right* is 0.42; that of *toe* and *finger* is 0.41; and that of *toe* and *left* is 0.06. Unlike distributional similarity, associativity is not directly rooted in language use (39). Therefore, while they are not fully independent measures either, they can diverge. For instance, *car* is distributionally similar to *bike* and associated with *petrol*. However, *bike* is not strongly associated with *car*, nor is *petrol* distributionally similar to it (40).

Finally, we also evaluate the first principal component (PC1) of associativity and distributional similarity. That is, we construct a measure of semantic relatedness out of these two sources of information, accounting for the largest amount of their variance. A priori, it is not clear whether this measure will characterize the data well.

Results. As shown in Table 1, cross-linguistic colexification patterns are best explained by the model with the PC1 measure of semantic relatedness. The ranking in the table, based on expected predictive accuracy, is only interpretable in relative terms, for model comparison. But the PC1 model also performs well in absolute terms: It has a root-mean-square error of 0.34, an accuracy of 0.84 when binarizing the mean of its posterior's predictions, and a Bayesian R^2 of 0.53 (41). For comparison, always predicting that data either does or does not colexify scores a root-mean-square error of 0.71 and an accuracy of 0.50.

	ELPD_Δ (SE_Δ)	ELPD (SE)	EFF (SE)
PC1	0.00 (0.00)	-77231.93 (266.11)	12.29 (0.19)
Associativity	-715.08 (366.14)	-77947.01 (266.16)	11.33 (0.21)
Distributional	-2145.77 (368.55)	-79377.70 (268.90)	12.64 (0.17)

Table 1. Model comparison of the PC1 model, the associativity model, and the distributional model using approximate leave-one-out cross-validation. ELPD_Δ is the difference in expected log point-wise predictive density to the best ranked model, PC1. EFF indicates the effective number of parameters.

Crucially, Figure 2 shows that the model identifies the hypothesized Goldilocks principle. It predicts that unrelated meanings, like THREE-YES, are unlikely to colexify. In line with previous research, as semantic relatedness increases, so does the likelihood to colexify (9). For instance, BRIGHT-YELLOW and TOWN-PEOPLE are more related than THREE-YES, and are thus more likely to be expressed by the same word in a language. This trend reaches an inflection point for highly related meanings, where the tendency reverses. For instance, TUESDAY-THURSDAY is the most related pair in the figure, and has a lower colexification likelihood than the less related pair CALF-CATTLE.

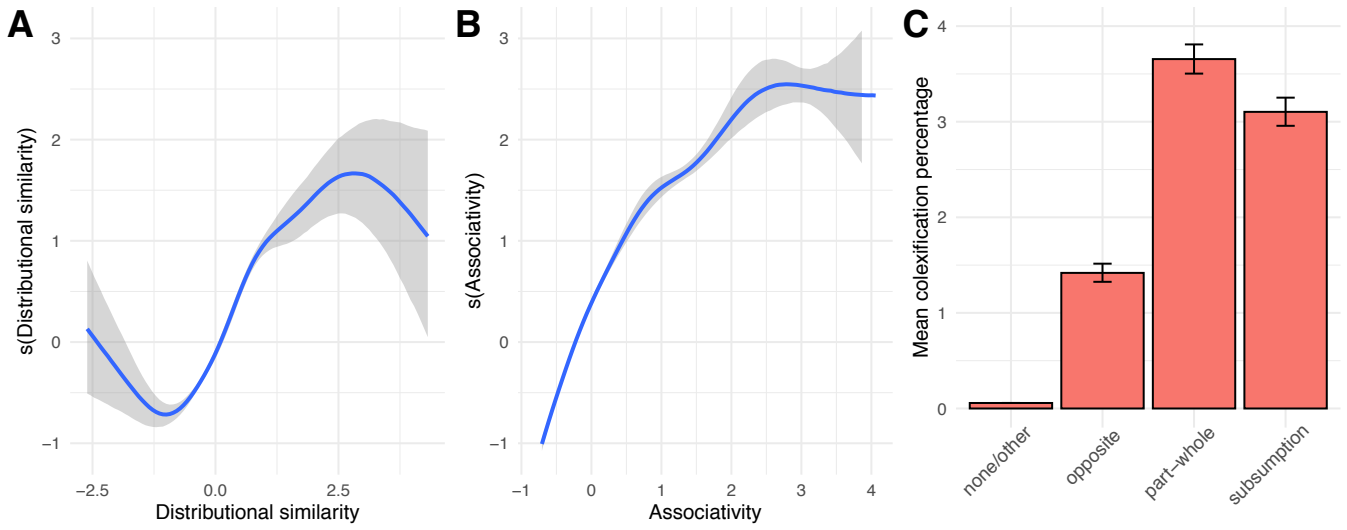

Fig. 3. A: Marginal effects of standardized distributional similarity. Shading shows 95% credible intervals. The smooth function $s(\cdot)$ is inferred from the data and characterizes how values on the x-axis contribute to colexification likelihood (on the logit scale). B: Marginal effects of standardized associativity. C: Mean percentage of colexification for meaning pairs, categorized by their semantic relations, with 95% credible intervals. With a region of practical equivalence of 1% (42), the part-whole and subsumption groups are equivalent in terms of their colexification rate, and all other groups differ from each other.

Confusability decreases colexification likelihood

The results so far show that high semantic relatedness can result in decreased colexification likelihood; however, our hypothesis specifically predicts that the shift in likelihood is due to **confusability**, rather than high semantic relatedness per se. We next provide two pieces of evidence that support the hypothesis, both focusing on the kind of relationship that meanings stand in, rather than the degree of relatedness.

First, the fact that distributional similarity, but not associativity, exhibits the Goldilocks curve (see Figure 3, panels A and B). What is more, both of the models that incorporate usage-based information –the distributional model and the PC1 model– identify the hypothesized curve. As noted above, high distributional similarity corresponds to high contextual overlap. Meanings that have such a large overlap in context are pressured to be expressed by different words, since otherwise they would not be distinguishable from one another (4, 27, 30, 31, 43). In other words, colexifying meanings that are hard to tease apart either risks misunderstandings or requires the speaker to provide further disambiguating material. In contrast, since associativity is not directly based on language use (39, 44, 45), we do not expect it to be related in the same way to confusability. And indeed, as shown in Figure 3B, for associativity we observe a plateau in colexification likelihood at the higher end of relatedness, but no decrease.

Second, if this explanation of the Goldilocks principle is on the right track, we would expect communicative pressure to make it less likely for languages to colexify opposite meanings (e.g., LEFT and RIGHT) than meanings related by other kinds of relationships, such as part-whole (TOE-FOOT). Opposite meanings express contrasts, being maximally similar in every respect but one (46–49). Therefore, losing the semantic distinction that they encode can be expected to be particularly harmful in communicative terms. Intuitions along these lines have been put forward in past studies (5, 9); we make a specific prediction, grounded in broader theoretical considerations, and test it empirically. As comparison points, we choose two semantic relations that do not necessarily lead to high confusability and can be estimated from existing resources (50): part-whole (e.g., TOE-FOOT) and subsumption (e.g., CALF-CATTLE; calves are cattle, therefore CATTLE subsumes CALF). Note that colexifying meanings connected by these relationships also still implies losing a potentially useful semantic distinction. However, we expect their rate of colexification to be higher than that of opposites, under the assumption that functional pressure exerts less force to lexically distinguish them.

Data. Colexification rates were estimated from 1416 meanings and 2279 languages from CLICS^3 (10). Semantic relations were extracted from WordNet (50), a human-annotated lexical database, using English words as proxies for meanings. Pairs that stand in none of the three examined relations were classified as 'none/other'. Details are given in *Material and Methods* and SI Section 1.D.

Results. Figure 3C shows mean colexification percentages for the different relationships. The results suggest, first, that standing in one of the three semantic relations increases the odds for meanings to colexify compared to the control group 'none/other'; and second, that not all relations are equally conducive to colexification. In particular, as predicted, meanings that stand in opposition to one another are less likely to be expressed by the same form than those standing in part-whole or subsumption relations. As in the preceding analysis, we thus find that semantic relatedness renders colexification more likely, but that the need to distinguish meanings that are particularly confusable can counteract this trend.

Discussion

We have found empirical support for a Goldilocks principle in colexification: Meanings are more likely to be expressed by the same word when they are neither too unrelated, nor too related, but just right. Specifically, our results suggest that the Goldilocks zone of colexification is composed of meanings

that are related enough that colexifying them fosters cognitive economy (9), and at the same time are not too confusable in actual language use. Our interpretation is that natural language lexicons follow the Goldilocks principle because they evolve to strike a balance between being as simple as possible while still being informative enough. That is, they do so as a response to competing cognitive and communicative pressures.

The pattern we identified is a tendency across languages; but we still expect important culture-specific effects on the way languages partition meanings into words (8). For instance, while languages tend to use different words for opposites, the meanings LEND and BORROW are still colexified in at least 40 languages. These languages are as phylogenetically and geographically varied as Thakali (Sino-Tibetan); Komi (Uralic); Guaraní (Tupian); and Takia (Austronesian). Also, as mentioned above, using the same word for two meanings that are related but not opposites, like TOE and FINGER, also implies losing a distinction that may be relevant for communicative success. Ultimately, while one linguistic community may not care to lexically distinguish LEND from BORROW, another may not care about keeping TOE and FINGER apart. In light of the diversity in how languages carve out reality through their lexicons, it is remarkable that a signature of the universal need to keep contextually confusable meanings apart can still be identified. We expect that using more and more diverse languages than Dutch and English to estimate semantic relatedness will shed further light on the patterns we have identified. We hope that the growing efforts to construct large-scale resources for minority languages will enable such an analysis in the future.

Our findings also have broader implications for phenomena regarding lexical ambiguity, in particular the pervasiveness of metaphor (51). Previous work (51, 52) indicates that it is common for metaphorically related senses to belong to different ontological domains, and, in particular, to vary along a concreteness-abstractness axis. As an example, the verb *go* in English can be used in a concrete physical sense ("Kids can easily go from the school to the library in this village") and in an abstract sense ("Voters can easily go from a liberal to a conservative position in this country"). It has furthermore been shown that metaphor is directional; for instance, historically, languages extend concrete words with abstract meanings (52). This has been suggested to be cognitively advantageous, because metaphor assists us in reasoning about abstract domains by extending features from domains that are more directly accessible to perception (51, 52). The present study suggests that metaphor is also advantageous from a functional perspective, because it allows speakers to conflate meanings without risking communicative failure: If two meanings belong to ontologically different domains, then it is unlikely that colexifying them will cause confusion in context. Under this interpretation, metaphor simultaneously maximizes simplicity and informativeness, which would explain its vast success as a linguistic mechanism. Future work should probe this hypothesis directly, and further examine how metaphor aids cognition (in particular, what specifically makes meanings relateable by metaphor), as well as how the hypothesis may extend to related semantic phenomena, such as metonymy.

More generally, this work contributes to the growing body of evidence that natural languages are shaped by the need for efficient communication, in the sense that they achieve a good balance between the two competing pressures for simplicity and informativeness (19–24). Going beyond the restricted domains examined so far (25, 26), the present study suggests that the trade-off between simplicity and informativeness is reflected in the way natural language lexicons associate words and meanings.

Materials and Methods

The data processing and analysis code developed for this article is available at: https://osf.io/hjvm5/. All the resources we use, cited below, are freely available.

Data pre-processing. The colexification data comes from CLICS^3 (10), a database that provides a standardized set of meanings and corresponding lexifications in over 2,000 languages. The geographic and phylogenetic information used in the first analysis was drawn from independent sources (53, 54). Geographic information (latitude and longitude of the place where each language is majoritarily spoken) was drawn from the Automated Similarity Judgment Program (53). Geographic distances are based on the shortest distance between two points on an ellipsoid. Estimated phylogenetic distances are from Jäger 2018 (54). They are based on the pointwise mutual information of word lists, and they have been shown to fare well at phylogenetic inference. Further details are given in SI Section 1.A.

For the first analysis, we excluded data from languages lacking phylogenetic or geographic information in our sources, as well as data for meanings that do not have Dutch or English lexifications in CLICS^3. We also excluded the data of the language providing the estimates for semantic relatedness (e.g., Dutch was left out of the analysis when it was used to obtain the relatedness estimates), to avoid circularity (9). We included all remaining positive cases of colexification as well as an equal number of negative examples, randomly sampled. We did not include all possible negative cases of colexification because it would make the analyses computationally intractable. This pre-processing strategy resulted in 203,056 data points encompassing 1453 unique meanings and 1259 distinct languages. SI Section 1 details all data pre-processing steps. SI Section 2 gives an overview of the resulting data sets.

Dutch and English distributional models are from fastText (55) (see SI Section 1.B for details). The associativity data comes from Small World of Words (38, 39). Following De Deyne et al. (39, 56), we consider three different transformations of the raw cue-response counts as measures of associativity. The measures are laid out in SI Section 1C. In the main text, we report results for the best one. Model comparison by means of differences in expected log point-wise predictive densities indicates that the best measure is the most sophisticated, random-walk based, transformation (see SI Table S3). This is in line with De Deyne et al.'s evaluations of these transformations on other semantic tasks (39).

Our second analysis, using WordNet (50), is based on 1416 meanings and 2279 languages (with a total of 1,001,742 data points). WordNet is a database of human-annotated semantic relations between lexical units. The primary WordNet unit is the so-called *synset*, or set of synonyms, aimed at representing a given sense of a word. A word can be included in different synsets. In our analysis, each meaning was represented by the most frequent synset of its English lexification in CLICS^3. The following semantic relations between synset pairs were then retrieved: antonymy (for opposite meanings), holonymy and meronymy (part-whole), and hyponymy and hypernymy (subsumption). The obtained data correspond to 79 antonyms, 70 holo-/meronyms, 155 hyper-/hyponyms, and 1,001,438 pairs that stand in none of these three relations. Details and descriptive statistics about the data drawn from WordNet are given in SI Section 1.D.

Models. Apart from a baseline intercept-only model, the generalized additive logistic regressions all have the general form

$$\text{logit}(p_{ijl}) = \beta_0 + \beta_1 \text{resource} + s(\text{rel}(i,j)) + \beta_2 P_{ijl} + \beta_3 G_{ijl}, \quad [1]$$

where the colexification of meanings i and j in language l is assumed to be Bernoulli distributed; *resource* indicates whether predictor information stems from Dutch or English; $\text{rel}(i,j)$ quantifies either

distributional similarity, or associativity, or PC1 of i and j, depending on the model; and P and G summarize how prevalent the colexification of i and j is in other languages k, weighted by the phylogenetic (P) or geographic (G) distance between l and k. Full model descriptions and comparisons are given in SI Section 3.

The general form of the distance variables P and G is

$$I_{ijl} \propto \sum\nolimits_k \Big(\text{colex}_{ijk} \;\; (1 - d(l,k)) \Big), \qquad [2]$$

with $\text{colex}_{ijk} = 1$ if meanings i and j colexify in language k and 0 otherwise; $k \neq l$; and $d(l,k)$ is the phylogenetic or geographic distance between l and k. P_{ijl} and G_{ijl} thus summarize how often meanings i and j are (not) colexified in languages other than l, factoring in their phylogenetic or geographic distance to l. Higher values indicate that two meanings are often colexified in neighboring languages. The converse is true for lower values. SI Section 3 provides definitions, and SI Table S2 shows that the formulation in (2) is preferrable to an exponentiated variant.

All models were diagnosed to ensure reliable estimates, and validated and compared using approximate leave-one-out cross-validation. Diagnostics and validations are reported in SI Section 3.A, comparisons in Section 3.B, and estimate summaries are given in Section 3.C.

ACKNOWLEDGMENTS. The authors thank Marco Baroni, Lucía Pitarch, and Igor Yanovich for their useful feedback and discussion. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 715154). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

1. GL Murphy, *The Big Book of Concepts.* (The MIT Press, Cambridge, MA), (2002).
2. T Wasow, A Perfors, D Beaver, The puzzle of ambiguity in *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe.* (CSLI Publications), pp. 265–282 (2005).
3. T Wasow, Ambiguity avoidance is overrated in *Ambiguity.* (Walter de Gruyter GmbH), (2015).
4. I Dautriche, Ph.D. thesis (École Normale Supérieure) (2015).
5. A François, Semantic maps and the typology of colexification: Intertwining polysemous networks across languages in *Studies in Language Companion Series.* (John Benjamins Publishing Company), pp. 163–215 (2008).
6. M Srinivasan, H Rabagliati, How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua* **157**, 124–152 (2015).
7. H Youn, et al., On the universal structure of human lexical semantics. *Proc. Natl. Acad. Sci.* **113**, 1766–1771 (2016).
8. JC Jackson, et al., Emotion semantics show both cultural variation and universal structure. *Science* **366**, 1517–1522 (2019).
9. Y Xu, K Duong, BC Malt, S Jiang, M Srinivasan, Conceptual relations predict colexification across languages. *Cognition* **201** (2020).
10. C Rzymski, et al., The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Sci. Data* **7**, 1–12 (2020).
11. S Kirby, JR Hurford, The emergence of linguistic structure: An overview of the iterated learning model in *Simulating the Evolution of Language*, eds. A Cangelosi, D Parisi. (Springer Verlag, London), pp. 121–148 (2002).
12. K Smith, S Kirby, H Brighton, Iterated learning: A framework for the emergence of language. *Artif. Life* **9**, 371–386 (2003).
13. S Kirby, T Griffiths, K Smith, Iterated learning and the evolution of language. *Curr. Opin. Neurobiol.* **28**, 108–114 (2014).
14. G Zipf, *Human behavior and the principle of least effort.* (Addison-Wesley Press), (1949).
15. A Martinet, *A Functional View of Language.* (Clarendon Press, Oxford), (1962).
16. LR Horn, Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature in *Meaning, Form and Use in Context*, ed. D Schiffrin. (Georgetown University Press), pp. 11 – 42 (1984).
17. G Jäger, R van Rooij, Language structure: psychological and social constraints. *Synthese* **159**, 99–130 (2007).
18. ST Piantadosi, Zipf's word frequency law in natural language: A critical review and future directions. *Psychon. bulletin & review* **21**, 1112–1130 (2014).
19. MH Christiansen, N Chater, Language as shaped by the brain. *Behav. Brain Sci.* **31** (2008).
20. T Regier, C Kemp, P Kay, *Word Meanings across Languages Support Efficient Communication.* (John Wiley & Sons, Ltd), pp. 237–263 (2015).
21. T Brochhagen, M Franke, R van Rooij, Coevolution of lexical meaning and pragmatic use. *Cogn. Sci.* **42**, 2757–2789 (2018).
22. S Kirby, M Tamariz, H Cornish, K Smith, Compression and communication in the cultural evolution of linguistic structure. *Cognition* **141**, 87–102 (2015).
23. E Gibson, et al., How efficiency shapes human language. *Trends Cogn. Sci.* **23**, 389–407 (2019).
24. JW Carr, K Smith, J Culbertson, S Kirby, Simplicity and informativeness in semantic category systems. *Cognition* **202**, 104289 (2020).
25. N Zaslavsky, C Kemp, T Regier, N Tishby, Efficient compression in color naming and its evolution. *Proc. Natl. Acad. Sci.* **115**, 7937–7942 (2018).
26. C Kemp, T Regier, Kinship categories across languages reflect general communicative principles. *Science* **336**, 1049–1054 (2012).
27. A Karjus, RA Blythe, S Kirby, T Wang, K Smith, Conceptual similarity and communicative need shape colexification: an experimental study (arXiv preprint).
28. C Ramiro, M Srinivasan, BC Malt, Y Xu, Algorithms in the historical emergence of word senses. *Proc. Natl. Acad. Sci. United States Am.* **115**, 2323–2328 (2018).
29. ST Piantadosi, H Tily, E Gibson, The communicative function of ambiguity in language. *Cognition* **122**, 280–291 (2012).
30. T Brochhagen, Signaling under uncertainty: Interpretative alignment without a common prior. *The Br. J. for Philos. Sci.* **71**, 471–496 (2020).
31. C Santana, Ambiguity in cooperative signaling. *Philos. Sci.* **81**, 398–422 (2014).
32. SN Wood, *Generalized Additive Models.* (Chapman and Hall/CRC), (2017).
33. M Westera, A Gupta, G Boleda, S Padó, Distributional models of category concepts based on names of category members. *Cogn. Sci.* **46** (to appear).
34. ZS Harris, Distributional structure. *Word* **10**, 146–162 (1954).
35. K Lund, C Burgess, Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. research methods, instruments, & computers* **28**, 203–208 (1996).
36. TK Landauer, ST Dumais, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211–240 (1997).
37. M Sahlgren, The distributional hypothesis. *Italian J. Disabil. Stud.* **20**, 33–53 (2008).
38. S De Deyne, DJ Navarro, G Storms, Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behav. Res. Methods* **45**, 480–498 (2013).
39. 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. Res. Methods* **51**, 987–1006 (2018).
40. F Hill, R Reichart, A Korhonen, SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Comput. Linguist.* **41**, 665–695 (2015).
41. A Gelman, B Goodrich, J Gabry, A Vehtari, R-squared for bayesian regression models. *The Am. Stat.* **73**, 307–309 (2019).
42. JK Kruschke, Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychol. Sci.* **6**, 299–312 (2011).
43. ST Piantadosi, H Tily, E Gibson, The communicative function of ambiguity in language. *Cognition* **122**, 280–291 (2012).
44. LB Szalay, J Deese, *Subjective meaning and culture: An assessment through word associations.* (Lawrence Erlbaum Hillsdale, NJ), (1978).
45. S Mollin, Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguist. Linguist. Theory* **5** (2009).
46. A Tversky, Features of similarity. *Psychol. Rev.* **84**, 327–352 (1977).
47. C Chiarello, C Burgess, L Richards, A Pollock, Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't ... sometimes, some places. *Brain Lang.* **38**, 75–104 (1990).
48. SM Mohammad, BJ Dorr, G Hirst, PD Turney, Computing lexical contrast. *Comput. Linguist.* **39**, 555–590 (2013).
49. T Kliegr, O Zamazal, Antonyms are similar: Towards paradigmatic association approach to rating similarity in SimLex-999 and WordSim-353. *Data & Knowl. Eng.* **115**, 174–193 (2018).
50. C Fellbaum, *WordNet.* (Oxford University Press), (2015).
51. G Lakoff, M Johnson, *Metaphors we live by.* (University of Chicago press), (1980).
52. Y Xu, BC Malt, M Srinivasan, Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cogn. Psychol.* **96**, 41–53 (2017).
53. S Wichmann, EW Holman, , CH Brown, The ASJP database (version 19) (2020).
54. G Jäger, Global-scale phylogenetic linguistic inference from lexical resources. *Sci. Data* **5** (2018).
55. E Grave, P Bojanowski, P Gupta, A Joulin, T Mikolov, Learning word vectors for 157 languages in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).* (2018).
56. S De Deyne, DJ Navarro, A Perfors, G Storms, Structure at every scale: A semantic network account of the similarities between unrelated concepts. *J. Exp. Psychol. Gen.* **145**, 1228–1254 (2016).

PNAS
www.pnas.org

Supplementary Information for

When do languages use the same word for different meanings? The Goldilocks Principle in colexification

Thomas Brochhagen and Gemma Boleda

Corresponding author: Thomas Brochhagen
E-mail: thomasbrochhagen@gmail.com

This PDF file includes:

Supporting Information Text

The following material details how the data from CLICS^3 (1) was processed and analyzed. Section 1 describes how the raw data was pre-processed and enriched. In particular, Section 1.A explains how phylogenetic and geographic information was added to it, and Sections 1.B through 1.D detail how distributional, associative, and relational information were obtained. Section 2 gives an overview of the resulting data sets, which our analyses take as starting points. Finally, Section 3 concerns the analyses reported in the main text: it specifies our models, lays out the diagnostics used to rule out issues with parameter estimates, and gives full numeric results for estimates and model comparisons.

Data availability. All the data and resources that our analyses build on are freely available. The CLICS^3 data (1) is available at https://clics.clld.org. FastText's (2) pre-trained word embeddings, at https://fasttext.cc/docs/en/crawl-vectors.html. The associativity data from *Small World of Words* (3, 4), at https://smallworldofwords.org/en/project. The phylogenetic distance information from (5), at https://osf.io/cufv7/. The geographic distance information from (6), at https://doi.org/10.5281/zenodo.3843469. WordNet (7), at https://wordnet.princeton.edu/.

Code availability. All data processing and analysis code used in this study is available at: https://osf.io/hjvm5/

1. Data preparation

The raw CLICS^3 data from (1) was pre-processed in the same way for all our analyses. We take a cross-linguistic, comparative perspective on what counts as distinct meanings. Two meanings are taken to colexify in a language if there is another language in which they are associated with two distinct words (1, 8, 9). We identify a language with its *glottocode* (10). This is a unique and stable identifier that most of the linguistic varieties found in CLICS^3 come associated with. We exclude data from languages lacking a glottocode; from two bookkeeping languoids (glottocodes *chua1256* and *bika1251*); and data from non-contemporary languages (e.g., Classical Greek or Middle English; 54 in total). The latter choice is motivated by not wanting to add a temporal dimension to our analysis. Lastly, we exclude glottocodes for which we lack phylogenetic or geographic information (see Section 1.A for details).

As detailed below, our analyses use three kinds of information to characterize the relationship in which meanings stand. The first is distributional information drawn from pre-trained neural network models (2). The second are associativity scores derived from human association data (3, 4). The third are semantic relationships, such as antonymy and hypernymy, extracted from the lexical database WordNet (7). Since meanings are not directly observable, the estimates for meaning relationships are all based on words in particular languages. For both distributional and associativity data, we use both Dutch and English resources. To extract specific semantic relationships like antonymy, the English WordNet is used. The Dutch WordNet was too small: we found no antonym pairs among our data.

Word-meaning associations from CLICS^3 were kept if there was a word form in the resource language (Dutch or English) expressing the relevant meaning available as well. In this way, for instance, the Dutch associativity data is used to calculate the association of all pairs of meanings for which CLICS^3 lists Dutch word forms. To avoid circularity and following (8), the word-meaning associations of the language providing the estimates are always removed. For example, colexification data from Dutch is removed when using Dutch associativity data.

For analyses that use distributional and associative information, the resulting data sets include all positive data points (meaning pairs that colexify) as well as an equal number of randomly sampled pairs of meanings that do not colexify. We do not include all possible negative cases of colexification because this makes these analyses computationally intractable. A meaning pair was considered to colexify in a language if at least one of the forms listed for the meanings in that language coincides (in CLICS^3, some meanings are linked to multiple forms in a language). Conversely, we consider a pair to not colexify in a language if the data for the language lists disjoint word forms for them. Each data point then consists of a meaning pair in a language together with its colexification status, distributional similarity score, associativity score, and indicators of phylogenetic and geographic distance (see below). The Dutch resource comprises 102,346 data points. It encompasses 1322 unique meanings and spans over 1245 distinct languages. The English resource contains 100,710 data points for 1274 unique meanings and 1253 languages.

For the analyses that use WordNet, all the meanings with English lexifications were used, including all positive and negative colexification cases. The resulting data set spans over 1416 unique meanings and 2279 unique languages, with 1,001,742 data points.

A. Phylogenetic and geographic distances. We enrich the pre-processed CLICS^3 data with geographic information from the Automated Similarity Judgment Program (ASJP) (6). More precisely, we add information about the latitude and longitude of where languages are majoritarily spoken. We draw on the ASJP for geographic information because the phylogenetic distances we employ are also based on it. Geographic distances were computed by estimating the shortest distance between two points on an ellipsoid, using the *distGeo* algorithm from the *geosphere* R package (see (11) for details). Distances were then rescaled in the range between 0 and 1 for ease of comparison with phylogenetic distances.

Estimates of phylogenetic distances between languages are from Jäger 2018 (5). They are based on the pointwise mutual information between ASJP word lists and range between 0 and 1. These estimates have been shown to fare well at phylogenetic inference when evaluated against the Glottolog expert tree (10). What is more, they are available for many of the languages found in CLICS^3, allowing us to enrich the data with distance information while retaining most of it.

B. Word embeddings. To compute distributional similarities between words, we use pre-trained word embeddings for English and Dutch that are available from the fastText library (2). We focus on Dutch and English because these are the same languages for which large scale associativity resources are also available. Additionally, the data that these models are trained on is large enough, and of high quality enough, to make them reliable (2). These word embeddings were obtained by running an extension of the CBOW algorithm (12) on data from Common Crawl and Wikipedia. They are optimized on a prediction task: predicting a target word given its context. The result is that word embeddings are similar if they appear in similar linguistic contexts.

C. Associativity measures. We use Dutch and English associativity data from Small World of Words (3, 4). We do so for two reasons. First, they are the largest associativity data sets to date for both languages. Second, they are openly available, and transparently collected and processed.

The English data from Small World of Words encompasses first, second, and third responses to 12282 cues (see (4) for details). The Dutch data covers 12571 cues (3). Responses to cues have been normalized, spell-checked and balanced such that each cue is judged by 100 subjects. To obtain measures of associativity from these data, we follow De Deyne et al. (4) and consider three transformations of Dutch and English cue-response counts: associative strength (Eq. 1), pointwise mutual information (Eq. 3), and a random walk-based measure (Eq. 5). In what follows, we give a concise description of each of these measures, focusing on the intuitions behind them. For further details and motivations see (4, 13).

These measures build on each other. The simplest is so-called *associative strength*. The associative strength between cues i and j is based on the cue-normalized number of times a response r is given for them:

$$\text{asso}_{str}(c_i, c_j) = \frac{\sum_k p(r_k \mid c_i)\, p(r_k \mid c_j)}{\sqrt{\sum_k p(r_k \mid c_i)^2}\sqrt{\sum_k p(r_k \mid c_j)^2}} \quad [1]$$

In words, associative strength reflects the amount and frequency by which responses are shared between two cues. It is a simple measure in the sense that it is only sensitive to the extent to which two cues share responses. In particular, associative strength does not consider how informative a response r is. One way to ameliorate this issue is to factor in the number of times a response r is given across cues:

$$\text{PPMI}(r \mid c) = \max\left(0, \log_2\left(\frac{p(r \mid c)}{\sum_i p(r \mid c_i) p(c_i)}\right)\right) \quad [2]$$

Exchanging the terms in (1) that codify the raw, frequency-based, relationship of a cue to a response by their pointwise mutual information in (2) then yields the second measure:

$$\text{asso}_{ppmi}(c_i, c_j) = \frac{\sum_k \text{PPMI}(r_k \mid c_i)\, \text{PPMI}(r_k \mid c_j)}{\sqrt{\sum_k \text{PPMI}(r_k \mid c_i)^2}\sqrt{\sum_k \text{PPMI}(r_k \mid c_j)^2}} \quad [3]$$

The third and final measure we consider is the decaying random walk-based transformation from De Deyne et al. (13). It not only considers direct links between two cues but also more indirect relationships. Intuitively, the relationship of two cues is given by the number of paths that lead from one to the other, and their length, in a network of cues. More technically, let $\mathbf{P}$ be a PPMI-transformed and normalized cue-response matrix. With $\mathbf{I}$ as the identity matrix, r as the maximum length of a walk, and a dampening parameter $\alpha < 1$ controlling the extent to which path-length plays a role, an associativity matrix $\mathbf{G}$ is obtained as follows (see (4, 13) for an explicit derivation):

$$\mathbf{G} = \sum_{r=0}^{\infty} (\alpha \mathbf{P})^r = (\mathbf{I} - \alpha \mathbf{P})^{-1} \quad [4]$$

As before, this information can then be plugged into (1), replacing $p(r \mid c)$ by the random walk-based value in $\mathbf{G}$ corresponding to the row of cue c and column of response r:

$$\text{asso}_{rw}(c_i, c_j) = \frac{\sum_k \mathbf{G}_{[r_k, c_i]}\, \mathbf{G}_{[r_k, c_j]}}{\sqrt{\sum_k (\mathbf{G}_{[r_k, c_i]})^2}\sqrt{\sum_k (\mathbf{G}_{[r_k, c_j]})^2}} \quad [5]$$

We compute the three measures (using $\alpha = 0.75$ for the random walk measure; see (4, 13)), on the full data sets with three responses per cue. In doing so, we follow De Deyne et al.'s observation that this yields a denser associativity network and, consequently, better estimates than if employing only first responses (4).

D. WordNet relations. WordNet (7) is a database of human-annotated semantic relations between sets of words. The primary WordNet unit is the so-called *synset*, or set of synonyms, aimed at representing a given sense of a word. For instance, the synset {*shiner*, *black eye*, *mouse*} represents the sense of these expressions that corresponds to "a swollen bruise caused by a blow to the eye" (see http://wordnetweb.princeton.edu/perl/webwn?s=mouse). WordNet defines the grouping of words into synsets and encodes semantic relationships between synsets.

We extracted data for the three semantic relationships of interest. Opposite meanings correspond to antonymy (e.g., *left* and *right* are antonyms); subsumption, to hypo-/hypernymy (*dog* is a hyponym of *animal*, and *animal* the hypernym of *dog*);

and part-whole, to holo-/meronymy (e.g., *foot* is a meronym of *leg*, and *leg* a holonym of *foot*). We take each pair of meanings i and j for which CLICS^3 lists English word forms, and extract the most frequent synsets s_i and s_j for the English words w_i and w_j lexifying i and j, respectively. We then register whether the synsets s_i and s_j stand in one of the three semantic relationships. If they do not, we record their relation as 'other/none' (meaning 'other semantic relationship or no semantic relationship listed in WordNet'). Pairs for which either w_i or w_j were not available in WordNet were discarded.

The obtained data correspond to 79 antonyms, 70 holo-/meronyms, 155 hyper-/hyponyms, and $1,001,438$ pairs that stand in none of the aforementioned relations. The most frequent colexification in this data set appears in 300 different languages and stands in a holo-/meronymy relation: LEG-FOOT. For this data set, we focus on colexification proportions: the number of languages in which a meaning pair colexifies divided by the total number of languages for which lexifications for both meanings are listed. This is to avoid effects driven by the mere frequency in which meanings are covered in CLICS^3.

2. Colexification data: descriptive overview

Figure S1 gives an overview of the data sets for associative and distributional information. It shows the relatedness scores for associative and distributional similarity across resources, split by whether a pair colexifies or not. The data are not aggregated; each data point represents a meaning pair i, j and a language l.

Table S1 reports estimated pairwise group differences in colexification percentage across WordNet relations, computed as explained in Section 1.D.

	none/other	holo-/meronymy	hyper-/hyponymy
antonymy	1.4 [1.3, 1.5]	-2.2 [-2.4, -2.1]	-1.7 [-1.9, -1.5]
none/other		-3.6 [-3.8, -3.4]	-3.1 [-3.2, -2.9]
holo-/meronymy			0.6 [0.3, 0.8]

Table S1. Pairwise group differences in colexification percentage, with 95% credible intervals. The 'none/other' group has the lowest mean percentage of colexification (0.06, with a 95% CI of [0.06, 0.06]), followed by antonyms (1.4 [1.3, 1.5]), then followed by hypo-/hypernyms and holo-/meronyms (3.1 [3.0, 3.3] and 3.7 [3.5, 3.8], respectively).

3. Models

All models are logistic regressions with the colexification of meanings i and j in language l as response:

$$\text{colex}_{ijl} \sim \text{Binomial}(1, p_{ijl}), \quad [6]$$

$$\text{colex}_{ijl} = \begin{cases} 1 \text{ if } i \text{ and } j \text{ colexify in } l, \\ 0 \text{ otherwise.} \end{cases} \quad [7]$$

The simplest model, (8), is an intercept-only model. It serves as a baseline for model comparison:

$$\text{logit}(p_{ijl}) = \beta_0 \quad [8]$$

All other models have three components in common: a categorical variable for the resource the predictors draw from (either English or Dutch); and two indicators for phylogenetic and geographic information, P_{ijl} and G_{ijl}. The first variable enables us to estimate an offset for distinct resources. The two indicators P_{ijl} and G_{ijl} allow us to factor in the effect that phylogenetic and geographic information have on the colexification of meanings i and j in language l. They summarize how prevalent the colexification of i and j is in other languages k, weighted by the phylogenetic or geographic distance between l and k. They are defined as:

$$P_{ijl} \propto \sum_k \Big(\text{colex}_{ijk} \;\; (1 - d_{\text{phy}}(l,k)) \Big), \quad [9]$$

$$G_{ijl} \propto \sum_k \Big(\text{colex}_{ijk} \;\; (1 - d_{\text{geo}}(l,k)) \Big), \quad [10]$$

with $k \neq l$; and $d_{\text{phy}}(l,k)$ and $d_{\text{geo}}(l,k)$ being the phylogenetic and geographic distances between l and k (Section 1.A). In words, P_{ijl} and G_{ijl} summarize how often meanings i and j are (not) colexified in other languages k, factoring in the phylogenetic or geographic distance between l and k. Higher values indicate that two meanings are often colexified in neighboring languages. Conversely, lower values indicate a lack of colexification of i and j in neighboring languages.

We also considered a different way to codify this information, exponentiating the distance information:

$$P^e_{ijl} \propto \sum_k \Big(\text{colex}_{ijk} \;\; \exp(1 - d_{\text{phy}}(l,k)) \Big), \quad [11]$$

$$G^e_{ijl} \propto \sum_k \Big(\text{colex}_{ijk} \;\; \exp(1 - d_{\text{geo}}(l,k)) \Big), \quad [12]$$

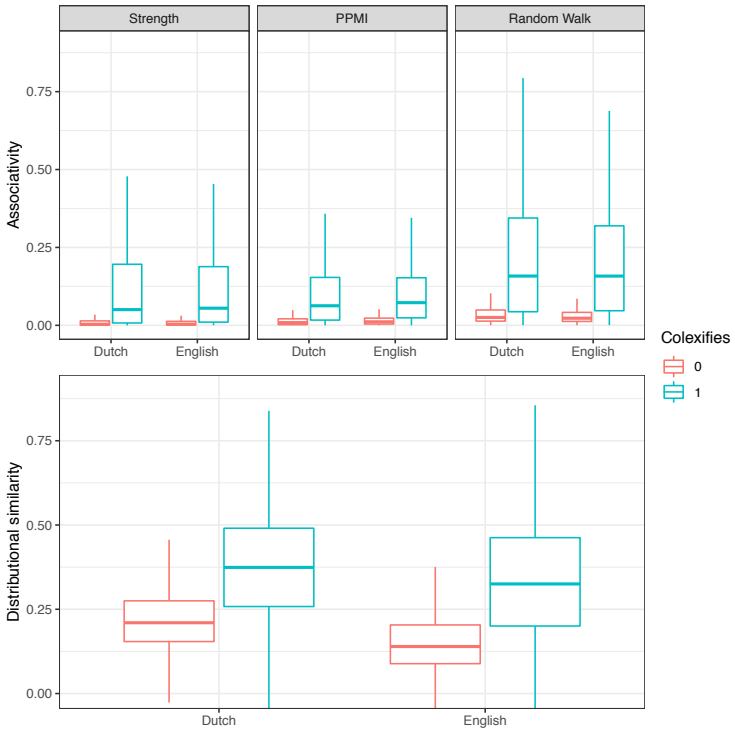

Fig. S1. Variation in the relationship of meaning pairs that do (not) colexify across different resources. Each panel represents a data set using the same measure. The top row shows associativity measures. The bottom row depicts distributional similarity. Each element on the x-axis is a distinct resource.

By contrast to (9) and (10), this puts more weight on information from nearby languages. As detailed in Section 3.B, no model does better with these exponentiated variants of P and G. This suggests a linear influence of distance on colexification (modulo the other predictors introduced below).

To summarize, apart from the intercept-only model in (8), all models are passed distance-weighted information about the pervasiveness of a colexification pattern in other languages, as well as the resource meta-language, which is either Dutch or English. This embodies the assumption that phylogenetic and geographic nearness can influence colexification (8, 14). Consequently, the models explicitly estimate both the effects of language contact and ancestry, as well as those independent of them.

The remaining models in (13-22) are generalized additive models. Each has one additional smooth term for semantic relatedness, estimated from either distributional information, associativity, or the first principal component of distributional information and random walk-based associativity (see Section 1 for information on how these measures were obtained). Finally, they differ on whether they use the unexponentiated distance indicators in (9) and (10); or their exponentiated variants in (11) and (12).

$$\text{logit}(p_{ijl}) = \beta_0 + \beta_1\ \text{resource} + s(\text{distr}(i,j)) \quad +\beta_2\ P_{ijl} + \beta_3\ G_{ijl} \quad [13]$$

$$\text{logit}(p_{ijl}) = \beta_0 + \beta_1\ \text{resource} + s(\text{asso}_{str}(i,j)) \quad +\beta_2\ P_{ijl} + \beta_3\ G_{ijl} \quad [14]$$

$$\text{logit}(p_{ijl}) = \beta_0 + \beta_1\ \text{resource} + s(\text{asso}_{ppmi}(i,j)) +\beta_2\ P_{ijl} + \beta_3\ G_{ijl} \quad [15]$$

$$\text{logit}(p_{ijl}) = \beta_0 + \beta_1\ \text{resource} + s(\text{asso}_{rw}(i,j)) \quad +\beta_2\ P_{ijl} + \beta_3\ G_{ijl} \quad [16]$$

$$\text{logit}(p_{ijl}) = \beta_0 + \beta_1\ \text{resource} + s(\text{pc1}(i,j)) \quad +\beta_2\ P_{ijl} + \beta_3\ G_{ijl} \quad [17]$$

$$\text{logit}(p_{ijl}) = \beta_0 + \beta_1\ \text{resource} + s(\text{distr}(i,j)) \quad +\beta_2\ P^e_{ijl} + \beta_3\ G^e_{ijl} \quad [18]$$

$$\text{logit}(p_{ijl}) = \beta_0 + \beta_1\ \text{resource} + s(\text{asso}_{str}(i,j)) \quad +\beta_2\ P^e_{ijl} + \beta_3\ G^e_{ijl} \quad [19]$$

$$\text{logit}(p_{ijl}) = \beta_0 + \beta_1\ \text{resource} + s(\text{asso}_{ppmi}(i,j)) +\beta_2\ P^e_{ijl} + \beta_3\ G^e_{ijl} \quad [20]$$

$$\text{logit}(p_{ijl}) = \beta_0 + \beta_1\ \text{resource} + s(\text{asso}_{rw}(i,j)) \quad +\beta_2\ P^e_{ijl} + \beta_3\ G^e_{ijl} \quad [21]$$

$$\text{logit}(p_{ijl}) = \beta_0 + \beta_1\ \text{resource} + s(\text{pc1}(i,j)) \quad +\beta_2\ P^e_{ijl} + \beta_3\ G^e_{ijl} \quad [22]$$

All models were fit in Stan (15) through the *brms* package (16). Each model was fit running 4 chains for 1000 iterations, with 500 iterations of warmup and default (non-informative) priors.

A. Diagnostics. All fits were checked to verify their estimates' reliability. None had divergent transitions nor saturated trajectory lengths. They all had an energy Bayesian Fraction of Missing information over 0.2 (17). They also all had parameters with a split $\hat{R} < 1.1$ (18). Finally, all fits had a large enough effective sampling size (> 0.001 effective samples per transition).

We use approximate leave-one-out cross-validation for validation and model comparison (19, 20). All cross-validations had a shape parameter $k < 0.7$. This suggests that the estimates are reliable.

B. Model comparisons. Table S2 shows pairwise comparisons between models with and without exponentiated distance terms. As advanced earlier, no model does better with exponentiated distance terms. Consequently, we consider only models with the unexponentiated distance terms (Eqs. (9) and (10)) in our main analyses.

Table S3 compares the three associativity models. In line with past comparisons of these three measures on other semantic tasks, the random walk-based measure outperforms its simpler alternatives (4).

Finally, Table S4 shows a comparison between the best associativity model, the distributional model, the intercept-only model, and the model with the first principal component of distributional and random walk-based associative information as a predictor (17). This last model provides the best fit.

C. Estimates. Table S5 gives a summary of the estimates for the best associativity model; the distributional model; and the model with their first principal component. Estimates of parameters that are common to all models are quite similar across model variations. The parameters encoding phylogenetic and geographic information show the same trends as in previous research (8, 14), with meanings being more likely to colexify in a language if they also colexify more often in languages that are phylogenetically or geographically near.

References

1. C Rzymski, et al., The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Sci. Data* **7**, 1–12 (2020).
2. E Grave, P Bojanowski, P Gupta, A Joulin, T Mikolov, Learning word vectors for 157 languages in *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. (2018).
3. S De Deyne, DJ Navarro, G Storms, Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behav. Res. Methods* **45**, 480–498 (2013).
4. 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. Res. Methods* **51**, 987–1006 (2018).
5. G Jäger, Global-scale phylogenetic linguistic inference from lexical resources. *Sci. Data* **5** (2018).
6. S Wichmann, EW Holman, , CH Brown, The ASJP database (version 19) (2020).

Table S2. Pairwise comparisons between models fit with exponentiated distance terms (labeled *exp*) and unexponentiated ones using approximate leave-one-out cross-validation. $ELPD_{\triangle}$ is the difference in expected log point-wise predictive density to the best ranked model. EFF indicates the effective number of parameters.

	$ELPD_{\triangle}$ ($SE_{\triangle}$)	ELPD (SE)	EFF (SE)
Associative strength	0.00 (0.00)	-80550.23 (268.25)	11.96 (0.40)
Associative strength (exp)	-1573.29 (233.60)	-82123.52 (265.47)	11.52 (0.34)
Associativity PPMI	0.00 (0.00)	-78352.19 (265.89)	11.59 (0.57)
Associativity PPMI (exp)	-1735.37 (227.51)	-80087.57 (263.88)	11.27 (0.58)
Associativity RW	0.00 (0.00)	-77947.01 (266.16)	11.33 (0.21)
Associativity RW (exp)	-1759.27 (225.98)	-79706.28 (264.13)	10.86 (0.19)
Distributional	0.00 (0.00)	-79377.70 (268.90)	12.64 (0.17)
Distributional (exp)	-1474.62 (236.03)	-80852.32 (266.52)	12.48 (0.16)
PC1	0.00 (0.00)	-77231.93 (266.11)	12.29 (0.19)
PC1 (exp)	-1857.46 (227.74)	-79089.38 (264.66)	11.57 (0.18)

Table S3. Model comparison of associativity models (14)-(16) using approximate leave-one-out cross-validation. $ELPD_{\triangle}$ is the difference in expected log point-wise predictive density to the best ranked model. EFF indicates the effective number of parameters.

	$ELPD_{\triangle}$ ($SE_{\triangle}$)	ELPD (SE)	EFF (SE)
Associativity RW	0.00 (0.00)	-77947.01 (266.16)	11.33 (0.21)
Associativity PPMI	-405.19 (42.68)	-78352.19 (265.89)	11.59 (0.57)
Associative strength	-2603.23 (76.78)	-80550.23 (268.25)	11.96 (0.40)

Table S4. Model comparison using approximate leave-one-out cross-validation. It compares the best associative model, the distributional model, the model using their first principal component, and the intercept-only model. $ELPD_{\triangle}$ is the difference in expected log point-wise predictive density to the best ranked model. EFF indicates the effective number of parameters.

	$ELPD_{\triangle}$ ($SE_{\triangle}$)	ELPD (SE)	EFF (SE)
PC1	0.00 (0.00)	-77231.93 (266.11)	12.29 (0.19)
Associativity RW	-715.08 (366.14)	-77947.01 (266.16)	11.33 (0.21)
Distributional	-2145.77 (368.55)	-79377.70 (268.90)	12.64 (0.17)
Intercept only	-63516.78 (266.11)	-140748.71 (0.01)	1.01 (0.00)

Table S5. Parameter estimates (posterior means) and standard errors with 95% credible intervals. The terms labeled as $sds(\cdot)$ are variance parameters. Larger values indicate wigglier smooths. Terms prefixed by $s.$ are the population-level parts of the spline.

Model		Estimate	Est.Error	l-95% CI	u-95% CI
PC1	Population-level effects				
	β_0	0.28	0.01	0.26	0.30
	β_2	1.37	0.02	1.33	1.41
	β_3	0.25	0.02	0.22	0.28
	$\text{resource}_{\text{english}}$	0.06	0.01	0.04	0.09
	s.PC1	-3.84	2.72	-9.22	1.12
	Smooth term				
	sds(PC1)	3.13	0.92	1.83	5.45
Associativity RW	Population-level effects				
	β_0	0.29	0.01	0.27	0.31
	β_2	1.43	0.02	1.39	1.47
	β_3	0.23	0.02	0.19	0.26
	$\text{resource}_{\text{english}}$	0.06	0.01	0.03	0.08
	$\text{s.asso}_{\text{rw}}$	10.69	2.38	6.23	15.25
	Smooth term				
	$\text{sds(asso}_{\text{rw}})$	2.38	0.84	1.13	4.36
Distributional	Population-level effects				
	β_0	0.23	0.01	0.21	0.25
	β_2	1.53	0.02	1.49	1.57
	β_3	0.26	0.02	0.22	0.29
	$\text{resource}_{\text{english}}$	0.08	0.01	0.05	0.10
	s.distr	-6.33	1.53	-9.42	-3.41
	Smooth term				
	sds(distr)	2.67	0.76	1.56	4.43

7. C Fellbaum, *WordNet*. (Oxford University Press), (2015).
8. Y Xu, K Duong, BC Malt, S Jiang, M Srinivasan, Conceptual relations predict colexification across languages. *Cognition* **201** (2020).
9. A Karjus, RA Blythe, S Kirby, T Wang, K Smith, Conceptual similarity and communicative need shape colexification: an experimental study (arXiv preprint).
10. H Hammarström, R Forkel, M Haspelmath, S Bank, Glottolog 4.3 (2020).
11. CFF Karney, Algorithms for geodesics. *J. Geod.* **87**, 43–55 (2012).
12. T Mikolov, K Chen, G Corrado, J Dean, Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
13. S De Deyne, DJ Navarro, A Perfors, G Storms, Structure at every scale: A semantic network account of the similarities between unrelated concepts. *J. Exp. Psychol. Gen.* **145**, 1228–1254 (2016).
14. JC Jackson, et al., Emotion semantics show both cultural variation and universal structure. *Science* **366**, 1517–1522 (2019).
15. B Carpenter, et al., Stan: A probabilistic programming language. *J. statistical software* **76** (2017).
16. PC Bürkner, brms: An R Package for Bayesian Multilevel Models using Stan. *J. Stat. Softw.* **80** (2017).
17. M Betancourt, A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434* (2017).
18. A Gelman, DB Rubin, Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
19. A Vehtari, A Gelman, J Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. computing* **27**, 1413–1432 (2017).
20. A Vehtari, J Gabry, M Magnusson, Y Yao, A Gelman, loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models (2019) R package version 2.2.0.