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

Thomas Brochhagen [a],[*], 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*

ARTICLE INFO

ABSTRACT

Lexical ambiguity is pervasive in language, and often systematic. For instance, the Spanish word *dedo* can refer to a toe or a finger, that is, these two meanings colexify in Spanish; and they do so as well in over one hundred other languages. Previous work shows that related meanings are more likely to colexify. This is attributed to cognitive pressure towards simplicity in language, as it makes lexicons easier to learn and use. The present study examines the interplay between this pressure and the competing pressure for languages to support accurate information transfer. We hypothesize that colexification follows a Goldilocks principle that balances the two pressures: meanings are more likely to attach to the same word when they are related to an optimal degree—neither too much, nor too little. We find support for this principle in data from over 1200 languages and 1400 meanings. Our results thus suggest that universal forces 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.

1. Introduction

The association of multiple meanings with the same form is pervasive across natural languages (Dautriche, 2015; Murphy, 2002; Wasow, 2015; Wasow et al., 2005), a phenomenon called **colexification** (François, 2008). For instance, as illustrated in Fig. 1A, the Spanish word *dedo* can refer to both a finger and a toe; that is, unlike English, Spanish *colexifies* these two meanings, using a single word to express both.[1] Many colexifications are attested throughout the world (François, 2008; Jackson et al., 2019; Srinivasan and Rabagliati, 2015; Xu et al., 2020a; Youn et al., 2016). For instance, the conflation of TOE and FINGER is found in at least 135 languages (Rzymski et al., 2020), many of which are phylogenetically unrelated and spoken in different parts of the globe. 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 data about colexification. The first force is 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 (a.o.,

Kirby and Hurford, 2002, Smith et al., 2003, Kirby et al., 2014). Regarding the lexicon, in the extreme, a very simple language could colexify all meanings, using a single word form to express them all. However, while very easy to learn and store, this language 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 (a.o., Zipf, 1949, Martinet, 1962, Horn, 1984, Jäger and van Rooij, 2007, Piantadosi, 2014, Christiansen and Chater, 2008, Regier, Kemp & Kay, 2015). At the other extreme, then, a maximally informative lexicon could have one distinct word per meaning, with no ambiguity. However, this would create larger lexicons that would be more difficult to learn and use: new meanings could not directly build on established word-meaning associations; and shared associations could not be exploited for the ease of lexical retrieval and interpretation (Ramiro et al., 2018; Srinivasan and Rabagliati, 2015; Xu et al., 2020a).

A growing body of research argues that languages are **efficient** in the sense that they strike a good balance between informativeness and simplicity (Brochhagen et al., 2018; Carr et al., 2020; Christiansen and Chater, 2008; Gibson et al., 2019; Kirby et al., 2015; Mollica et al., 2021;

---

* Corresponding author.
  E-mail address: thomas.brochhagen@upf.edu (T. Brochhagen).

[1] Colexification refers to the same phenomenon as ambiguity. However, the two notions take different perspectives: If a word is ambiguous (e.g. *dedo*), then the meanings it expresses (e.g. TOE and FINGER) colexify. Colexification encompasses both polysemy and homonymy.

Regier et al., 2015). Some of this work considers the lexicon (Mollica et al., 2021; Regier et al., 2015); however, so far only restricted domains have been explored. These include color (Zaslavsky et al., 2018); numerals (Xu et al., 2020b); quantifiers (Steinert-Threlkeld, 2021); containers (Xu et al., 2016); indefinites (Denić et al., 2021); and kinship (Kemp and Regier, 2012). In the present study, we examine the interaction between simplicity and informativeness in the lexicon at a broader 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 (Karjus et al., 2021; Xu et al., 2020a). This tendency has been attributed to cognitive 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 (Ramiro et al., 2018; Srinivasan and Rabagliati, 2015; Xu et al., 2020a). However, in line with Karjus et al.'s (2021) findings using artificial languages, we hypothesize that informativeness may counterbalance the tendency to colexify related meanings: If meanings are too related, then expressing them with the same form can be disadvantageous from a functional, communicative point of view. For instance, LEFT and RIGHT are highly related but are often relevant alternatives in context. Consider someone giving directions; if they say *go left*, there is often the contextually relevant alternative of going right. Thus, using the same form for LEFT and RIGHT risks leading to communicative failure. Indeed, the possibility to contextually disambiguate meanings is crucial for the persistence of lexical ambiguity (Brochhagen, 2020; Piantadosi et al., 2012; Santana, 2014).

Note that it is always possible to disambiguate meanings using longer expressions; for instance, Spanish speakers can use *dedo del pie* ('finger of the foot') when they need to unambiguously refer to a toe. Analogously, it would also be possible to use a single word, for instance *dax*, for LEFT and RIGHT, and to use a more complex expression to distinguish between the two. What is at stake is thus not whether languages can express a given semantic distinction, but whether they care enough about it to encode it in the lexicon. The prediction is that, on average, they will care more about distinctions that are often alternatives in context, because the communicative need to distinguish them is higher, with context providing less information to tease them apart.

To sum up, we expect the communicative need to distinguish meanings to play a role in shaping lexicons across languages. Communicative need varies across language communities depending on factors such as environment and culture (Jackson et al., 2019; Kemp et al., 2018; Xu et al., 2020a). However, we predict that the pressure for informativeness will show a universal signature over and above such language-specific variation.

More concretely, 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". The Goldilocks principle is illustrated in Fig. 1B. Crucially, following the hypothesis that what hinders communication is meaning confusability in context (Brochhagen, 2020; Piantadosi et al., 2012), we expect "too related" to mean "too confusable". In other words, we expect colexification likelihood to decrease for highly related meanings where confusability is at stake. As discussed above, this particularly concerns sets of meanings that are contrasting alternatives to each other. Examples of such meanings are weekdays such as MONDAY and TUESDAY, meanings related to quantification like SOME and ALL, and opposites like WARM and COLD.

We find support for the Goldilocks principle in two analyses. The first uses data-induced measures of semantic relatedness to characterize 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. However, a decrease in likelihood is only partially confirmed: while the data is best characterized by a decrease, it is also consistent with a plateau, suggesting that informativeness may exert less force than we expected a priori. The second analysis further probes the role of confusability in the shift in colexification likelihood for related meanings. We find that meanings that are often alternatives in context, in particular those that express opposites, are indeed less likely to colexify than other kinds of related meanings. Our results thus support the hypothesis that natural language lexicons evolve to strike a balance between competing pressures for simplicity and informativeness.

2. Colexification follows the Goldilocks principle

To study the relationship between semantic relatedness and colexification, we fit regression models to colexification data.[2] The data comes from CLICS³ (Rzymski et al., 2020), the largest cross-linguistic database of colexifications available to date. This database is the result of a standarized aggregation of multiple typological datasets, e.g., the Intercontinental Dictionary Series (Key & Comrie, 2021) and NorthEuraLex (Dellert and Jäger, 2017). This is accomplished by interfacing with other resources such as Glottolog (Hammarström et al., 2020) – for the unification of information about the language varieties involved – and Concepticon (List et al., 2016) – providing comparative meaning glosses. The Concepticon catalogue, in turn, is the outcome of an aggregation and unification of concepts from multiple meaning list datasets.[3] All in all, CLICS³ provides a standardized set of meanings and corresponding lexifications in over 3000 languages. In what follows, two distinct meanings are taken to colexify if they share a lexification, i.e., if they are expressed by the same word in the database.

In this first analysis, we proceed in two steps: we first identify the operationalization of the variables of interest that best explains the data. This is independent of the shape of the effects; the best model could or could not show the Goldilocks curve. Once we find the best model, we inspect its estimate of the effect of semantic relatedness on colexification.

2.1. Models

We use generalized additive logistic models (Wood, 2017), which allow for non-linear relationships between the dependent variable and independent variables. This makes them suitable to probe our hypothesized relationship between colexification likelihood and semantic relatedness (Fig. 1B). Generalized additive models include a penalization against excessive curvature: "wigglier" trends, such as the Goldilocks curve compared to a (more) linear relationship, are only established if they substantially improve the model fit.[4]

The models characterize how likely a 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. Since language contact – facilitated by geographic proximity – and common linguistic ancestry influence colexification (Jackson et al., 2019; Xu et al., 2020a), 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

---

[2] 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.

[3] Future work may benefit from the NoRaRe dataset (Tjuka et al., 2021), which maps the Concepticon concepts used in CLICS³ to word and concept properties in several languages.

[4] Notwithstanding, for explicitness' sake, linear versions of the models reported on in the main text are compared to their (possibly) non-linear counterparts in SI Section 3.2. In all cases, additive models outrank their linear counterparts.
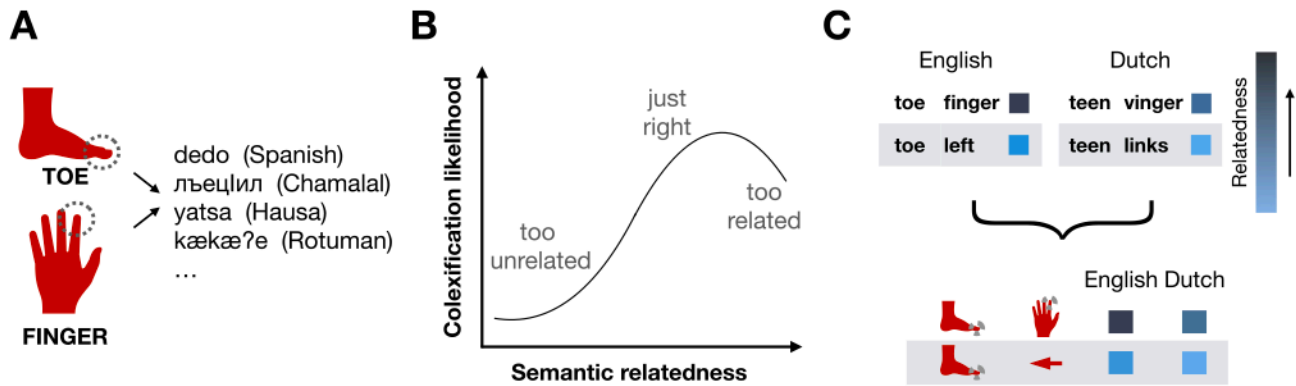
Fig. 1. A: Illustration of cross-linguistic systematicity in colexification. B: Hypothesized relationship between the relatedness of meanings and the likelihood that they are expressed by the same form. The extremes on the x-axis are areas where meanings are either too related or too unrelated to be likely to be colexified. Unrelated meanings (e.g., FINGER and KETTLE) 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. Measures of semantic relatedness, such as distributional similarity, are computed on word pairs and used for meaning pairs. For instance, the distributional similarity of the words *teen* and *vinger* in Dutch (upper right part of the figure) serves as a proxy for the similarity of the meanings TOE and FINGER (lower part of the figure). The distributional similarity of the corresponding English words is taken as an alternative estimate of these meanings' relatedness.

response language. More precisely, all models 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 resources (see below); rel(i,j) is a data-induced estimate of semantic relatedness; 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. The smooth function $s(\cdot)$ corresponds to the potentially non-linear contribution of relatedness, rel(i,j), on colexification likelihood (Wood, 2017).[5]

The general form of the distance variables P and G is

$$I_{ijl} \propto \sum_k \left(\text{colex}_{ijk} \quad (1 - d(l,k)) \right), \quad (2)$$

with colex$_{ijk} = 1$ if meanings i and j colexify in language k and 0 otherwise; $k \neq l$; and $d(l,k)$ being the phylogenetic or geographic distance between l and k. P_{ijl} and G_{ijl} thus summarize how often meanings i and j are 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.

Geographic information – latitude and longitude of the place where each language is majoritarily spoken – was drawn from Glottolog (Hammarström et al., 2020), and provided through CLICS[3]. Geographic distances are based on the shortest distance between two points on an ellipsoid. Identifying a language with a single point on the globe is a clear simplification, particularly for linguistic communities spanning large regions; and inaccurate for languages spoken in different parts of the world (e.g., English or Spanish). Consequently, while both issues are strongly mitigated by the fact that they are comparably rare in the large sample of languages we analyse, they can lead to noisy estimates for

some individual languages. Phylogenetic distance estimates are from Jäger (2018). They are based on the pointwise mutual information of word lists. These estimates have been shown to fare well at phylogenetic inference. Further details and discussion on distance information are given in SI Section 1.1.

All models were diagnosed to ensure reliable estimates, and validated and compared using approximate leave-one-out cross-validation (Vehtari et al., 2017, 2019). Individual model definitions are given in full in SI Section 3: diagnostics and validations are reported in Section 3.1, comparisons in Section 3.2, and estimate summaries in Section 3.3. SI Table S2 shows that the formulation of distance indices as in (2) is preferable to an exponentiated variant.

Pre-processing the data from CLICS[3] for this first analysis yielded 203,056 data points, encompassing 1453 unique meanings and 1259 distinct languages. This includes all positive cases of colexification from the database for which we had information conforming with Eq. (1) as well as an equal number of negative examples, randomly sampled. We did not include all possible negative cases of colexification because that would make the analyses computationally intractable. SI Section 1 details all pre-processing steps and SI Section 2 gives an overview of the resulting data sets.

2.2. Estimating semantic relatedness

We follow previous work in using words as surrogates for meanings when estimating semantic relatedness (e.g., Karjus et al., 2021; Westera, Gupta, Boleda, & Padó, 2021; Xu et al., 2020a). More specifically, we use words in Dutch and English (previous work used English only). As illustrated in Fig. 1C, the relatedness of word pairs, such as *teen-vinger* in Dutch or the equivalent *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 meanings in other languages (Eq. 1). It would be desirable to use more – and more linguistically diverse – languages to estimate semantic relatedness; however, at present only Dutch and English have resources that are large enough, and of a high enough quality, for our analysis. SI Section 1 discusses this issue in more detail.

Building on Xu et al. (2020a), we evaluate three measures of

---

[5] While a maximally random structure would be desirable, adding random intercepts or slopes makes the models computationally intractable on a cluster with 500GB of RAM. We thus decided to trade off model structure in favor of data coverage and, chiefly, in favor of the inclusion of a non-parametric form for the relationship of relatedness to colexification.

semantic relatedness: distributional similarity, associativity, and the first principal component of these two measures.[6]

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 (Harris, 1954; Landauer and Dumais, 1997; Lund and Burgess, 1996). The Dutch and English distributional models that we use are from fastText (Grave et al., 2018). To illustrate the measure, the contexts of use of *left* and *right* are quite similar (distributional similarity of 0.57 in the English model, with 1 being the maximum); *toe* and *finger* are also quite similar but less so (0.47); and *toe* and *bird* are, expectedly, the least similar of these pairs (0.06).

Associativity is derived from large-scale association norms from De Deyne et al. (2013, 2018), 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*. Following De Deyne et al. (2016, 2018), we consider three different transformations of the raw cue-response counts as measures of associativity. The measures are laid out in SI Section 1.3. 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 this is the most sophisticated, random-walk based, transformation (see SI Table S3). This is consistent with De Deyne et al.'s (2018) evaluation of these transformations on other semantic tasks. Using the examples from above and the English associativity scores that we use in this study, *left* and *right* have an associativity of 0.42 (maximum is 1); *toe* and *finger* score 0.41; and *toe* and *bird* score 0.02.

Distributional similarity and associativity codify different facets of semantic relatedness, but they do not strongly diverge either. They have a Pearson's correlation of 0.7 for Dutch resources; 0.82 for English resources; and 0.76 overall. To intuitively exemplify where they may differ: *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 (Hill et al., 2015). This motivates the use of a third measure that synthesizes the two "views" on semantic relatedness given by distributional similarity and associativity, namely, their first principal component (PC1). PC1 accounts for the largest amount of the variance of the two measures. A priori, it is not clear how well PC1 will characterize the data. If both distributional similarity and associativity are relevant to colexification in complementary ways, then it is likely that their first principal component will be as well – and possibly even more so. Conversely, if, instead, either distributional similarity or associativity is starkly less informative about colexification than the other measure, then the synthesis provided by PC1 will also be less successful than the more informative measure it is based on. SI Section 2 gives a visual overview of the colexification data in relation to the different measures of relatedness employed.

2.3. Results

Table 1 shows a comparison of the three operationalizations of semantic relatedness as predictors of colexification. It shows that cross-linguistic patterns are best explained by the model with the PC1 measure of semantic relatedness. Thus, distributional similarity and associativity provide complementary information about the kind of relatedness that matters for colexification. The ranking in Table 1, based on expected predictive accuracy, is only interpretable in relative terms, for model comparison. However, the PC1 model also performs well in

Table 1
Model comparison of the PC1 model, the associativity model, and the distributional model using approximate leave-one-out cross-validation (Vehtari et al., 2017). All three are generalized additive models that have the colexification of a pair of meanings in a language as dependent variable and one of three operationalizations of semantic relatedness as an independent variable (Eq. 1). ELPD_Δ is the difference in expected log point-wise predictive density to the best ranked model, PC1. Intuitively, ELPD evaluates a model against an estimate of future data, weighted by how likely this data is estimated to be. EFF indicates the effective number of parameters. It serves as an indicator of a model's complexity. The three models are approximately equivalent in this respect.

	ELPD_Δ (SE_Δ)	ELPD (SE)	EFF (SE)
PC1	0.00 (0.00)	−77,231.93 (266.11)	12.29 (0.19)
Associativity	−715.08 (366.14)	−77,947.01 (266.16)	11.33 (0.21)
Distributional	−2,145.77 (368.55)	−79,377.70 (268.90)	12.64 (0.17)

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 (Gelman et al., 2019). For comparison, a random baseline model would obtain a root-mean-square error of 0.71 and an accuracy of 0.50.

We next turn to the main hypothesis. Fig. 2 shows that the best model identifies the hypothesized Goldilocks principle. The left graph in the figure depicts the marginal effect of semantic relatedness, and the right part shows model predictions for example meaning pairs.[7] The model estimates that unrelated meanings, like THREE-YES, are unlikely to colexify. In line with previous research, for low to medium values, as semantic relatedness increases, so does the likelihood to colexify (Xu et al., 2020a). 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. However, as hypothesized, this trend breaks for highly related meanings. For instance, TUESDAY-THURSDAY is the most related pair in the figure, and has a lower mean colexification likelihood than the less related pair CALF-CATTLE.

As shown in Fig. 2.A, the data is most compatible with a decrease in colexification likelihood at the higher end of semantic relatedness (see blue line). However, it is also compatible with a plateau (see upper part of shaded area). Either way, the model identifies a clear shift in regime, with a non-linear relationship between semantic relatedness and colexification likelihood. The data thus support the hypothesis that, for highly related meanings, the positive relationship between semantic relatedness and colexification likelihood does not hold anymore. We return to this matter in Section 4.

3. Confusability decreases colexification likelihood

The results so far suggest that there is a shift in colexification likelihood for highly related meanings; however, our hypothesis specifically predicts that the shift is due to **confusability**, rather than high semantic relatedness per se. We next probe the role of confusability directly.

As discussed above, we expect communicative pressure to make it less likely for languages to colexify meanings that often express contrasting alternatives to each other in context. In Fig. 2B, this is exemplified by the pairs NORTH-SOUTH, STALLION-MARE and THURSDAY-THURSDAY. The notion of contextually relevant alternative is intuitively clear and relevant to many areas of linguistics, but to the best of our knowledge no independent definition of it exists (see Buccola et al., 2021 for further discussion). For this reason, we focus on opposites (e.g., LEFT and RIGHT), a subset of such contextually relevant alternatives for which independent

---

[6] Xu et al. (2020a) additionally consider frequency and two variables related to metaphoricity. These factors were found to be less informative about colexification than distributional similarity and associativity. SI Section 3 shows results for models with frequency added as an additional predictor. They indicate that the effects reported below are neither explained nor modulated by frequency.

---

[7] For completeness' sake, the marginal effects of distributional similarity and associativity are depicted in SI Section 3.3. However, it is important to stress that the PC1 model best characterizes the colexification data and thus provides the most reliable estimate of the relationship between semantic relatedness and colexification likelihood that we have at our disposal.
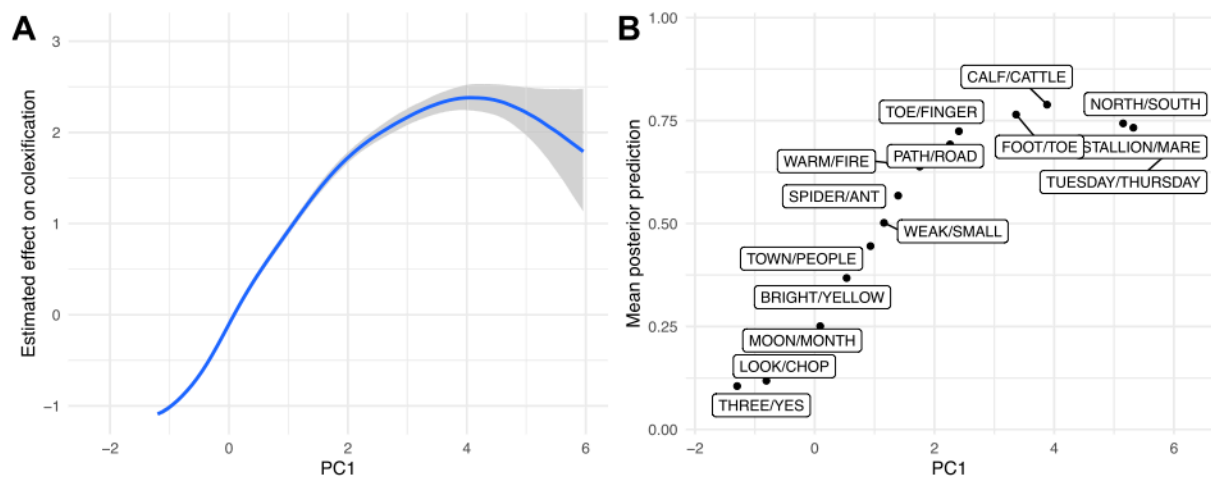
Fig. 2. A: Marginal effects of the best measure of semantic relatedness (PC1, in standardized units). Shading shows 95% credible intervals. A smooth function is inferred from the data and characterizes how the contribution of PC1 to colexification likelihood changes across its values (on the logit scale); this is depicted on the *y*-axis. Uncertainty increases with deviation from the predictor's mean. This is expected given that data in this region is comparatively sparse (see SI Section 2 for an overview of the data distribution). 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.

resources exist (Fellbaum, 2015).

Opposite meanings express contrasts, being maximally similar in every respect but one (Chiarello et al., 1990; Kliegr and Zamazal, 2018; Mohammad et al., 2013; Tversky, 1977). 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 (François, 2008; Xu et al., 2020a); we here make a specific prediction, grounded in broader theoretical considerations, and probe it empirically. As comparison points, we choose two semantic relations that do not necessarily lead to high confusability and can also be estimated from existing resources (Fellbaum, 2015): 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 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.

For this analysis, colexification rates for the different semantic relations were estimated from 1416 meanings and 2279 languages from CLICS[3] (Rzymski et al., 2020). Semantic relations were extracted from WordNet (Fellbaum, 2015), a human-annotated lexical database, using English words as proxies for meanings. 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 this 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. Data not covered by WordNet was not included in the analysis. Further details and descriptive statistics are given in SI Section 1.5.

3.1. Results

Fig. 3 shows mean colexification percentages for the different relationships. These 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, thus, we find that semantic relatedness renders colexification more likely; and, moreover, we show that the need to distinguish meanings that are particularly confusable can counteract this trend. In our interpretation, thus, simplicity pushes the colexification rate for opposites up, and informativeness pulls it down, resulting in the middle position of opposites (with respect to the other semantic relations) shown in Fig. 3.

A further piece of evidence that contextual confusability may be at play is the fact that opposites have a higher mean distributional similarity (0.59, SD −0.18) than meanings in the part-whole (0.46, SD
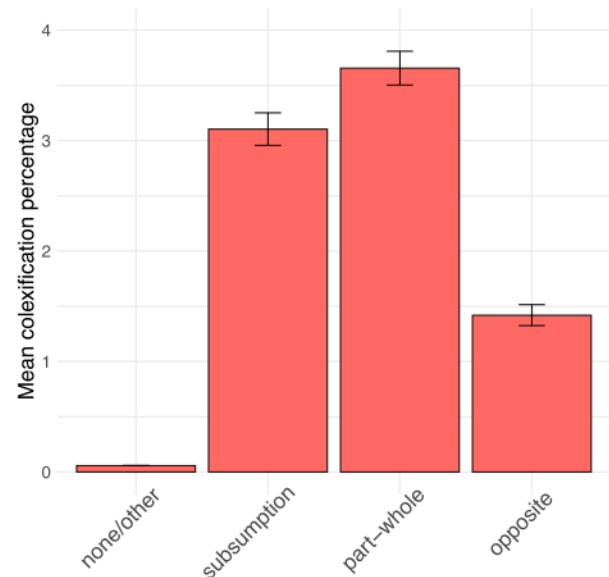

Fig. 3. Mean colexification percentage for meaning pairs, categorized by semantic relations, with 95% credible intervals. With a region of practical equivalence of 1% (Kruschke, 2011), part-whole and subsumption groups are equivalent in terms of colexification rates; all other groups differ from each other.

=0.14) and subsumption (0.42, SD =0.16) relations.[8] This indicates that the contexts of use of opposites are more similar (recall that distributional similarity captures contextual overlap). Therefore, one can expect it to be harder – on average – to tell opposites apart, rendering them less likely to colexify, due to pressure for informativeness.

4. 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. This pattern is predicted by the measure of semantic relatedness that best characterizes the colexification data: a synthesis of distributional similarity and associativity. More specifically, our results suggest that the Goldilocks zone of colexification is composed of meanings that are related enough that colexifying them fosters cognitive economy (Karjus et al., 2021; Xu et al., 2020a), and at the same time are not too confusable in actual language use (Karjus et al., 2021). 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.

We should stress that, while less likely according to model estimates, the results of the first analysis do not rule out a weaker version of the principle, one in which colexification likelihood does not decrease with high semantic relatedness, but simply plateaus. This version still supports the general hypothesis, namely, that pressure for informativeness counteracts the increasing trend from simplicity. However, it also suggests that informativeness may exert less force than we expected a priori. Another caveat is the fact that we have used a particular database of colexifications (Rzymski et al., 2020); while this is the most complete source of colexification data available to date, it could be that it leads to underestimating colexification rates for particular kinds of meanings, or to other kinds of biases in the results. In particular, the database covers just under 3000 concepts, most of them pertaining to concrete objects that are commonly relevant in language communities; and only certain subsets of these concepts are covered in each of the languages included in the database. Future work should aim at an even broader coverage of the conceptual domain and the world's languages.

While the pattern we identified is a tendency across languages, we still expect important culture-specific effects on the way languages partition meanings into words, depending on their communicative needs (Jackson et al., 2019; Xu et al., 2020a). For instance, while languages tend to use different words for opposites, the meanings LEND and BORROW are still colexified in at least 40 languages (Rzymski et al., 2020). 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 (Kemp et al., 2018; Xu et al., 2020a). 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 be identified.

Throughout this study we focused on the relationship of *pairs* of meanings to better understand what drives some of them together. This contrasts with the kind of characterization provided by comparative studies on semantic maps (e.g., Croft, 2001; Haspelmath, 1997; Haspelmath, 2003), where the relations between meanings are mapped out in a network-like structure to uncover universal implicational patterns

like "if a form in a language expresses both x and y, it also expresses z" (see also François, 2008, List et al., 2013 and Jackson et al., 2019 for network-based studies of colexification). To the best of our knowledge, while being similar in scope and aims, these two approaches have not yet been integrated. A promising open area of research would consequently be to extend the kind of analysis conducted here to networks, and elucidate how and to which extent implicational patterns can be derived from them.

Our findings have broader implications for phenomena regarding lexical ambiguity, in particular the pervasiveness of metaphor (Lakoff and Johnson, 1980). Previous work (Lakoff and Johnson, 1980; Xu et al., 2017) 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 (Xu et al., 2017). 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 (Lakoff and Johnson, 1980, Xu et al., 2017). Our 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, we contribute 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 (Brochhagen et al., 2018; Carr et al., 2020; Christiansen and Chater, 2008; Dingemanse et al., 2015; Gibson et al., 2019; Kirby et al., 2015; Monaghan et al., 2011; Regier et al., 2015). Going beyond the restricted domains examined so far (Denić et al., 2021; Kemp and Regier, 2012; Steinert-Threlkeld, 2021; Xu et al., 2016; Xu et al., 2020b; Zaslavsky et al., 2018), our work suggests that the trade-off between simplicity and informativeness is reflected in the way natural language lexicons associate words and meanings, and how they manage ambiguity, shedding further light on how universal principles shape language.

CRediT authorship contribution statement

Thomas Brochhagen: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Gemma Boleda:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Funding acquisition.

---

[8] This result is in line with previous work in distributional semantics (e.g. Lin et al., 2003).

Acknowledgments

The authors thank Marco Baroni, Lucía Pitarch, Igor Yanovich, Yang Xu and Dan Dediu for their useful feedback and discussion.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105179.

References

Brochhagen, T. (2020). Signaling under uncertainty: Interpretative alignment without a common prior. *The British Journal for the Philosophy of Science, 71*(2), 471–496. https://doi.org/10.1093/bjps/axx058

Brochhagen, T., Franke, M., & van Rooij, R. (2018). Coevolution of lexical meaning and pragmatic use. *Cognitive Science, 42*(8), 2757–2789. https://doi.org/10.1111/cogs.12681

Buccola, B., Križ, M., & Chemla, E. (2021). Conceptual alternatives. *Linguistics and Philosophy.* https://doi.org/10.1007/s10988-021-09327-w

Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition, 202,* 104289. https://doi.org/10.1016/j.cognition.2020.104289. URL doi:10.1016%2Fj.cognition.2020.104289.

Chiarello, C., Burgess, C., Richards, L., & Pollock, A. (1990). Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't… sometimes, some places. *Brain and Language, 38*(1), 75–104. https://doi.org/10.1016/0093-934X(90)90103-N. ISSN 0093-934X https://www.sciencedirect.com/science/article/pii/0093934X9090103N.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences, 31*(05). https://doi.org/10.1017/s0140525x08004998

Croft, W. (2001). *Radical construction grammar.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198299554.001.0001. oct. URL doi:10.1093%2Facprof%3Aoso%2F9780198299554.001.0001.

Dautriche, I. (2015). *Weaving an ambiguous lexicon.* PhD thesis, École Normale Supérieure.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2018). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods, 51*(3), 987–1006. https://doi.org/10.3758/s13428-018-1115-7. URL doi:10.3758%2Fs13428-018-1115-7.

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General, 145*(9), 1228–1254. https://doi.org/10.1037/xge0000192. . URL doi:10.1037%2Fxge0000192.

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods, 45*(2), 480–498. https://doi.org/10.3758/s13428-012-0260-7. URL doi:10.3758%2Fs13428-012-0260-7.

Dellert, J., & Jäger, G. (Eds.). (2017). *NorthEuraLex (version 0.9).* Max Planck Institute for Evolutionary Anthropology, Tübingen: Eberhard-Karls University. URL https://github.com/lexibank/northeuralex/releases.

Denić, M., Steinert-Threlkeld, S., & Szymanik, J. (2021). Complexity/informativeness trade-off in the domain of indefinite pronouns. *Semantics and Linguistic Theory, 30,* 166. https://doi.org/10.3765/salt.v30i0.4811. URL doi:10.3765%2Fsalt.v30i0.4811.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences, 19*(10), 603–615.

Fellbaum, C. (2015). *WordNet.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199842193.013.001. apr. URL doi:10.1093%2Foxfordhb%2F9780199842193.013.001.

François, A. (2008). Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In *Studies in language companion series* (pp. 163–215). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.106.09fra.

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for Bayesian regression models. *The American Statistician, 73*(3), 307–309. may https://doi.org/10.1080/00031305.2018.1549100. URL doi:10.1080%2F00031305.2018.1549100.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences, 23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003. ISSN 1364-6613.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the international conference on language resources and evaluation (LREC 2018).*

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2020). *Glottolog 4.3.* https://doi.org/10.5281/zenodo.4061162. URL.

Harris, Z. S. (1954). Distributional structure. *Word, 10*(2–3), 146–162.

Haspelmath, M. (1997). *Indefinite pronouns.* Oxford University Press.

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In *The new psychology of language* (pp. 217–248). Psychology Press. https://doi.org/10.4324/9781410606921-11. jan. URL.

Hill, F., Reichart, R., & Korhonen, A. (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics, 41*(4), 665–695. dec https://doi.org/10.1162/coli_a_00237. URL doi:10.1162%2Fcoli_a_00237.

Horn, L. R. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form and use in context* (pp. 11–42). Georgetown University Press.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., … Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science, 366*(6472), 1517–1522. https://doi.org/10.1126/science.aaw8160

Jäger, G. (2018). Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data, 5*(1). https://doi.org/10.1038/sdata.2018.189. URL doi:10.1038%2Fsdata.2018.189.

Jäger, G., & van Rooij, R. (2007). Language structure: Psychological and social constraints. *Synthese, 159*(1), 99–130. https://doi.org/10.1007/s11229-006-9073-5

Karjus, A., Blythe, R. A., Kirby, S., Wang, T., & Smith, K. (2021). Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science, 45*(9). https://doi.org/10.1111/cogs.13035. sep. URL doi:10.1111%2Fcogs.13035.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science, 336*(6084), 1049–1054. https://doi.org/10.1126/science.1218811

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics, 4*(1), 109–128. jan https://doi.org/10.1146/annurev-linguistics-011817-045406. URL doi:10.1146%2Fannurev-linguistics-011817-045406.

Key, M. R., & Comrie, B. (Eds.). (2021). *The Intercontinental Dictionary Series.* Leipzig: Max Planck Institute for Evolutionary Anthropology. URL https://ids.clld.org/.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology, 28,* 108–114. https://doi.org/10.1016/j.conb.2014.07.014

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi, & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 121–148). London: Springer Verlag.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition, 141,* 87–102. https://doi.org/10.1016/j.cognition.2015.03.016

Kliegr, T., & Zamazal, O. (2018). Antonyms are similar: Towards paradigmatic association approach to rating similarity in SimLex-999 and WordSim-353. *Data & Knowledge Engineering, 115,* 174–193. may https://doi.org/10.1016/j.datak.2018.03.004. URL doi:10.1016%2Fj.datak.2018.03.004.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299–312. may https://doi.org/10.1177/1745691611406925. URL doi:10.1177%2F1745691611406925.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* University of Chicago Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240. https://doi.org/10.1037/0033-295X.104.2.211. URL doi:10.1037/0033-295x.104.2.211.

Lin, D., Zhao, S., Qin, L., & Zhou, M. (2003). Identifying synonyms among distributionally similar words. *IJCAI, 3,* 1492–1493. Citeseer.

List, J.-M., Cysouw, M., & Forkel, R. (2016). Concepticon: A resource for the linking of concept lists. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 2393–2400).

List, J.-M., Terhalle, A., & Urban, M. (2013). Using network approaches to enhance the analysis of cross-linguistic polysemies. In *Proceedings of the 10th international conference on computational semantics - Short papers* (pp. 347–353). Stroudsburg.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203–208.

Martinet, A. (1962). *A functional view of language.* Oxford: Clarendon Press.

Mohammad, S. M., Dorr, B. J., Hirst, G., & Turney, P. D. (2013). Computing lexical contrast. *Computational Linguistics, 39*(3), 555–590.

Mollica, F., Bacon, G., Zaslavsky, N., Xu, Y., Regier, T., & Kemp, C. (2021). The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences, 118*(49). https://doi.org/10.1073/pnas.2025993118. e2025993118, dec. URL doi:10.1073%2Fpnas.2025993118.

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General, 140*(3), 325.

Murphy, G. L. (2002). *The big book of concepts.* Cambridge, MA: The MIT Press.

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review, 21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition, 122*(3), 280–291. https://doi.org/10.1016/j.cognition.2011.10.004

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences of the United States of America, 115,* 2323–2328. jun 10.31234/osf.io/8scm9. URL doi:10.31234%2Fosf.io%2F8scm9.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. *The Handbook of Language Emergence* (pp. 237–263). John Wiley & Sons, Ltd.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., , … Kaiping, G. A., et al. (2020). The database of cross-linguistic

colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data, 7* (1), 1–12.

Santana, C. (2014). Ambiguity in cooperative signaling. *Philosophy of Science, 81*(3), 398–422. https://doi.org/10.1086/676652

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life, 9*, 371–386.

Srinivasan, M., & Rabagliati, H. (2015). How concepts and conventions structure the lexicon: Cross-linguistic evidence from polysemy. *Lingua, 157*, 124–152. https://doi.org/10.1016/j.lingua.2014.12.004

Steinert-Threlkeld, S. (2021). Quantifiers in natural language: Efficient communication and degrees of semantic universals. *Entropy, 23*(10), 1335. Oct. 10.3390/e23101335.URL doi:10.3390%2Fe23101335.

Tjuka, A., Forkel, R., & List, J.-M. (2021). Linking norms, ratings, and relations of words and concepts across multiple language varieties. *Behavior Research Methods*, 1–21.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352. https://doi.org/10.1037/0033-295x.84.4.327. URL doi:10.1037%2F0033-295x.84.4.327.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., & Gelman, A. (2019). loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models. URL https://mc-stan.org/loo.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432.

Wasow, T. (2015). Ambiguity avoidance is overrated. In *Ambiguity*. Walter de Gruyter GmbH. https://doi.org/10.1515/9783110403589-003.

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. In *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe* (pp. 265–282). CSLI Publications.

Westera, W., Gupta, A., Boleda, G., & Padó, S. (2021). Distributional models of category concepts based on names of category members. *Cognitive Science, 46*.

Wood, S. N. (2017). *Generalized additive models*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279. may. URL doi:10.1201%2F9781315370279.

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020a). Conceptual relations predict colexification across languages. *Cognition, 201*.

Xu, Y., Liu, E., & Regier, T. (2020b). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind, 4*, 57–70. aug https://doi.org/10.1162/opmi_a_00034. URL doi:10.1162%2Fopmi_a_00034.

Xu, Y., Malt, B. C., & Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive Psychology, 96*, 41–53. https://doi.org/10.1016/j.cogpsych.2017.05.005. URL doi:10.1016%2Fj.cogpsych.2017.05.005.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science, 40*(8), 2081–2094. oct https://doi.org/10.1111/cogs.12312. URL doi:10.1111%2Fcogs.12312.

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., … Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences, 113*(7), 1766–1771. https://doi.org/10.1073/pnas.1520752113. URL doi:10.1073%2Fpnas.1520752113.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences, 115*(31), 7937–7942. jul https://doi.org/10.1073/pnas.1800521115. URL doi:10.1073%2Fpnas.1800521115.

Zipf, G. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.