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

Keywords:
Language universals
Colecification
Cognitive effort
Ambiguity
Efficient communication

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. 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 τός and τοιασί 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;
Regier et al., 2015). Some of this work considers the lexicon (Molllica 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 (Karijus 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 Karijus 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 particular operationalization 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. 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 ontologies. All in all, CLICS provides a standarized set of meanings and corresponding lexicifications in over 3000 languages. In what follows, two distinct meanings are taken to colexify if they share a lexicification, 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 fingers 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

³ 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.

⁴ Future work may benefit from the NoRaRe dataset (Tjuka et al., 2021), which maps the Concepticon concepts used in CLICS to word and concept properties in several languages.

⁵ Notwithstanding, for explicitness’ sake, linear versions of the models reported 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.
response language. More precisely, all models have the general form
\[
\logit(p_{ijl}) = \beta_0 + \beta_{\text{resource}} + s(\text{rel}(i,j)) + \beta_2 P_{i,j} + \beta_3 G_{i,j},
\]
where the colexification of meanings \(i\) and \(j\) in language \(l\) is assumed to be Bernoulli distributed; \(\text{resource}\) indicates whether predictor information stems from Dutch or English resources (see below); \(\text{rel}(i,j)\) 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, \(\text{rel}(i,j)\), on colexification likelihood (Wood, 2017).\(^5\)

The general form of the distance variables \(P\) and \(G\) is
\[
l_{ijl} \propto \sum_{k} \text{colex}_{ik} \left(1 - d(l,k)\right),
\]
with \(\text{colex}_{ik} = 1\) 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 \(k\). 
\(P_{i,j}\) and \(G_{i,j}\) 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\(^5\). 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 Jager (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 (Vehhtari 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 preferrable to an exponentiated variant.

Pre-processing the data from CLICS\(^5\) 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; Western, Gupta, Boleda, & Pado, 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
semantic relatedness: distributional similarity, associativity, and the first principal component of these two measures.

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 S2). 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 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 R² 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. 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. 2A, 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

|        | ELPD₁ (SE) | ELPD₁ (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) |

Note: ELPD₁ 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.

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.
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 (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, 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 lexicification in CLICS. 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...
This result is in line with previous work in distributional semantics (e.g. Lin et al., 2003).
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105179.
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](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](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. [https://doi.org/10.3390/e23101335](https://doi.org/10.3390/e23101335).

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](https://doi.org/10.1037/0033-295x.84.4.327).

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.

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