Semantic chaining and efficient communication: The case of container names

Yang Xu (yang_xu_ch@berkeley.edu) Terry Regier (terry.regier@berkeley.edu)
Department of Linguistics
University of California, Berkeley

Barbara C. Malt (barbara.malt@lehigh.edu)
Department of Psychology
Lehigh University

Abstract

Semantic categories in the world's languages often reflect a historical process of *chaining*: A name for one idea is extended to a conceptually related idea, and from there on to other ideas, producing a chain of concepts that all bear the same name. The beginning and end points of such a chain might in principle be conceptually rather dissimilar. There is also evidence supporting a contrasting picture: Languages tend to support efficient, informative communication, often through semantic categories in which all exemplars are similar. Here, we explore this tension through computational analyses of existing cross-language naming and sorting data from the domain of household containers. We find: (1) formal evidence for historical semantic chaining, and (2) evidence that systems of categories in this domain nonetheless support near-optimally efficient communication. Our results suggest that semantic chaining may be constrained by the functional need for efficient, informative communication.

Keywords: semantic variation; artifact categories; semantic chaining; historical semantics; semantic universals; efficient communication

Introduction

Languages vary widely in the ways they partition human experience into categories. For example, some languages use a single color term to cover both green and blue (Berlin & Kay, 1969), and some languages have spatial terms that highlight notions such as "attachment by spiking" (Levinson & Meira, 2003)—a notion that is not captured in the basic spatial lexicon of English. Yet at the same time, many logically possible semantic categories are not attested, and similar categories appear in unrelated languages. What explains this pattern of wide but constrained variation?

An existing proposal holds that this variation may be explained by the functional need for *efficient communication*—that is, the need to communicate precisely, using minimal cognitive resources. On this account, the different category systems that we see across languages constitute different means to this same functional end. This idea has accounted for cross-language variation in semantic domains including color (Regier, Kay, & Khetarpal, 2007), kinship (Kemp & Regier, 2012), space (Khetarpal, Neveu, Majid, Michael, & Regier, 2013) and number (Xu & Regier, 2014). It also coheres naturally with a recent focus on efficient communication as an explanation for other aspects of language (Piantadosi, Tily, & Gibson, 2011; Fedzechkina, Jaeger, & Newport, 2012; Smith, Tamariz, & Kirby, 2013). Importantly for our present purposes, in several of the above studies of semantic categories, efficient communication is shown to be supported by tightly-clustered coherent categories in which all exemplars tend to be similar to each other.

This proposal appears to conflict with a well-established and influential idea: that semantic categories reflect a historical process of *chaining*, whereby a name for one idea is extended to a related idea, and then on to further ideas, resulting in a chained structure in which the later items in the chain may have little similarity to the early ones (Lakoff, 1987; Brugman, 1988; Heit, 1992; Bybee, Perkins, & Pagliuca, 1994; Hopper & Traugott, 2003). For example, it has been suggested that such semantic chaining over historical time may explain the extensions of English container names such as *bottle* and *jar*. Malt, Sloman, Gennari, Shi and Wang (1999) found that the extensions for such container names include exemplars that are dissimilar to exemplars within the category on average, but are very similar to certain individual exemplars, consistent with the idea of chaining. Sloman, Malt and Fridman (2001) found that a computational model that captures chain-like structures accounted well for the English naming data. These analyses examined the data without reference to historical information, and thus did not directly assess whether the data are consistent with chaining over historical time. But they do appear to challenge the proposal that semantic systems support communicative efficiency through categories in which exemplars tend to be similar to each other.

Two important questions are left open by this earlier work. First, is there evidence for a genuinely historical process of chaining in container naming? And second, if so, does chaining in this domain in fact hamper efficient communication? Or is this semantic domain, like others, shaped by the need for efficient communication, despite semantic chaining? The studies we present address these questions.

In what follows, we summarize the theory of efficient communication, and demonstrate that chaining is in principle a challenge to this theory. We also briefly describe the cross-language data on which we rely. We then present two studies based on those data. The first study tests for historical chaining in the naming of containers, and the second study tests whether container naming across languages is communicatively efficient. To preview our results, we find evidence for historical chaining, yet we also find that despite this chaining, the container naming systems of three languages all support near-optimally informative communication. We conclude that semantic chaining may be constrained by the need for categories to be informative.

Formal presentation of theory

In this section, we present the theory of efficient communication in formal terms, following the formulation of Regier et al. (2015). We then demonstrate that chaining can lead to inefficient communication.

Consider the communicative scenario of Figure 1. Here, the speaker has a target object in mind—in this case, a specific kind of bottle—and wishes to communicate that idea to the listener. To that end, the speaker utters the word *bottle*. Given that utterance, the listener then attempts to *mentally reconstruct* the speaker's intended meaning. Because the word *bottle* covers a range of possible objects, the listener's representation is inexact and is shown as a probability distribution extending over that range. We take a communicative system to to be *informative* to the extent that it supports accurate mental reconstruction by the listener of the speaker's intended meaning; that is, reconstruction that is as exact as possible.

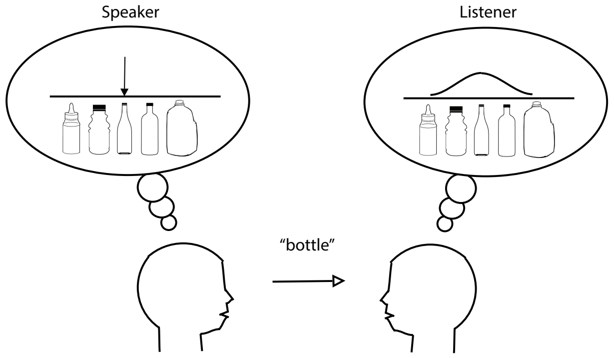

Figure 1: A simple communicative scenario.

We model the mental representations of both speaker and listener as probability distributions. Unlike the listener's distribution, the speaker's distribution S is certain: it consists of a point mass centered on the target, capturing our assumption that the speaker is certain of the meaning she wishes to convey. Following Regier et al.'s (2015) analysis of color naming, we take the listener distribution $L(i)$ to be based on the similarity (assessed empirically) of exemplar i to all exemplars in the category named by the word w:

$$L(i) \propto \sum_{j \in w} sim(i, j) \tag{1}$$

We then take the unit *communicative cost* $C(i)$ of communicating about a target object i using a particular communicative system to be a measure of the discrepancy between the listener distribution L and the speaker distribution S: specifically the Kullback-Leibler divergence $D_{KL}(S||L)$ between these two distributions. In the case of speaker certainty, this reduces to surprisal:

$$C(i) = D_{KL}(S||L) = \sum_j S(j) \log_2 \frac{S(j)}{L(j)} = \log_2 \frac{1}{L(i)} \tag{2}$$

Finally, we take the overall communicative cost of a system to be the expected communicative cost incurred in communicating about the domain. This is the sum of the unit costs of all possible targets in the domain, each weighted by its relative frequency of occurrence in usage, or need probability $N(i)$:

$$E[C] = \sum_i C(i)N(i) \tag{3}$$

We take a communicative system to be *informative* to the extent that it exhibits low communicative cost $E[C]$. We take the *complexity* of a system to be the number of lexical categories in the system. Finally, we take a system of categories to be *communicatively efficient* to the extent that it it is more informative than most logically possible hypothetical systems of the same complexity.

Chaining and inefficient communication

Semantic chaining can give rise to inefficient communication, as illustrated in Figure 2.

Panel (a) of this figure shows two artificial category systems that partition the same set of 8 objects (shown as black dots). The category system in the left half of the panel divides these objects into two non-chained (or clustered) categories. The system in the right half of the panel divides the same set of objects into two chained categories. The complexity of the two systems is the same (2 categories in each system), but they differ in informativeness. Panel (b) shows the communicative costs of these two systems, compared with the costs of all hypothetical partitions of the 8 objects into 2 groups of size 4.[1] It can be seen that relative to these hypothetical

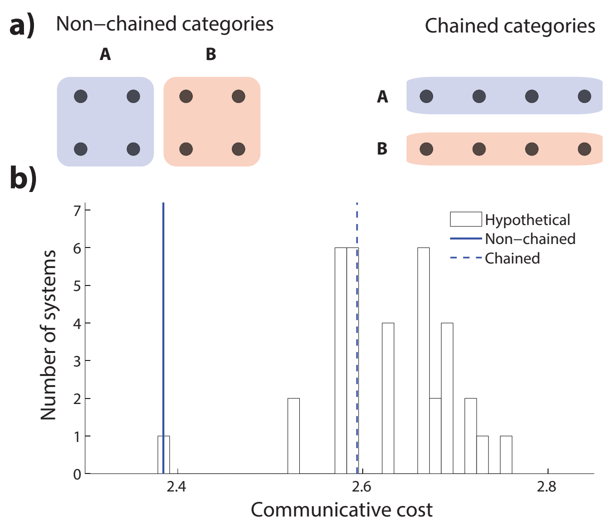

Figure 2: Chaining and communicative inefficiency. a) Non-chained and chained systems of equal complexity (2 categories each). b) Communicative costs of these systems and hypothetical systems of equal complexity.

[1]In computing cost, we assumed that the distance between horizontally or vertically neighboring objects in the grid is 1,

systems, the non-chained system from panel (a) is optimally informative for this level of complexity, and thus communicatively efficient, whereas the chained system is not. This demonstrates that semantic chaining has the potential to yield inefficient communication, as formalized here.

Data

We reanalyze data from Malt et al. (1999) on the naming and perceived similarity of household containers. The stimulus set for this research consisted of photographs of 60 household containers, representing a wide range of bottles, jars, and similar containers. Figure 3 shows a sample of these objects.

Figure 3: Sample stimuli from Malt et al. (1999).

We used two types of data that had been collected relative to this stimulus set. The first type is linguistic. Native speakers of American English, Mandarin Chinese, and Argentinian Spanish were instructed to name each container stimulus in their native language, giving whatever name they felt was best.[2] The second type of data is from a pile-sorting task. We used pile-sorting by English and Chinese speakers from the study by Malt et al. (1999) (for which data were retrievable), and we focused on sorting based on overall similarity of the containers (i.e. both the physical appearances and functions of the containers). We aggregated pile-sorting responses from subjects across languages, and took the similarity $sim(i, j)$ of any two objects i and j to be the proportion of all participants who sorted those two objects into the same pile. These naming and similarity measures were used in our analyses below.

Study 1: Historical chaining

In our first study, we asked whether these data provide evidence for historical chaining. That is, has the current extension of the names been developed through a chain of uses expanding over historical time? We test for historical chaining over all three languages in the dataset.

We considered three common categorization models from the literature, specified in Table 1: a chaining model, a clustering model, and a majority vote model. The chaining model is a nearest-neighbor (or 1-nearest-neighbor) model, which assigns a target item to the category that includes the exemplar most similar to that target item; this is the model that was explored by Sloman et al. (2001). The clustering model is based on Equation 1, and assigns a target item to the category whose examplars exhibit the greatest similarity to the target overall. The majority vote model is a baseline model that assigns a target item to the category that has the most exemplars, without reference to any intrinsic relations among exemplars. Similarities $sim(i, j)$ were determined by the pile-sort data, and the category (word) w for each container object was determined by the modal head noun that was used to label it in the naming data.[3]

Table 1: Summary of models. In the rules below, i is the target exemplar, j is any exemplar other than i, w is a lexical category, $sim(\cdot,\cdot)$ is the similarity between two exemplars, and $|w|$ is the size of category w.

Model	**Categorization rule**
Chaining	$i \rightarrow$ the category w of arg $\max_j sim(i, j)$
Clustering	$i \rightarrow$ arg $\max_w \sum_{j \in w} sim(i, j)$
Majority vote	$i \rightarrow$ arg $\max_w |w|$

Each model was tested against the data in a manner that recapitulates the addition of new exemplars to categories over historical time. We began by time-stamping the name of each container item, providing an estimate of when that item appeared in history, as follows. For each container item in the dataset, we determined the modal modifier- and head-noun phrase for that item (e.g. *juice bottle*) from the naming task, and performed a corpus search for that phrase in a large historical corpus, the Google Ngram American English corpus (Michel et al., 2011), over the period 1800-2000, and recorded the frequency of use of that phrase for each year. For each container phrase, we then applied a change-point detection algorithm (Kass, Eden, & Brown, 2014, sections 14.2.1 and 14.2.2) to these historical frequency traces, to determine the year in which each phrase experienced a substantial rise in frequency from a baseline of zero; we took that year to be the date of emergence of that object with respect to the head noun. We then simulated the sequential emergence of these exemplars in history, and asked which of the three models specified above best accounted for categorizations found in the naming data.

For each model, this predictive analysis proceeded as follows. We first seeded the model with the earliest item that appeared in history. As each remaining item became avail-

that the similarity between any two objects i, j is $sim(i, j) = \exp(-distance(i, j))$, and that need probability $n(i)$ is uniform across all objects i.

[2]Naming data were collected from 28 native speakers of English, all students at Lehigh University in the United States, 51 native speakers of Spanish, all from Comahue National University in Argentina, and 50 native speakers of Mandarin Chinese, 10 of whom were students at Lehigh and 40 of whom were students at Shanghai University in China.

[3]E.g. the category for *juice bottle* would be *bottle* in English.

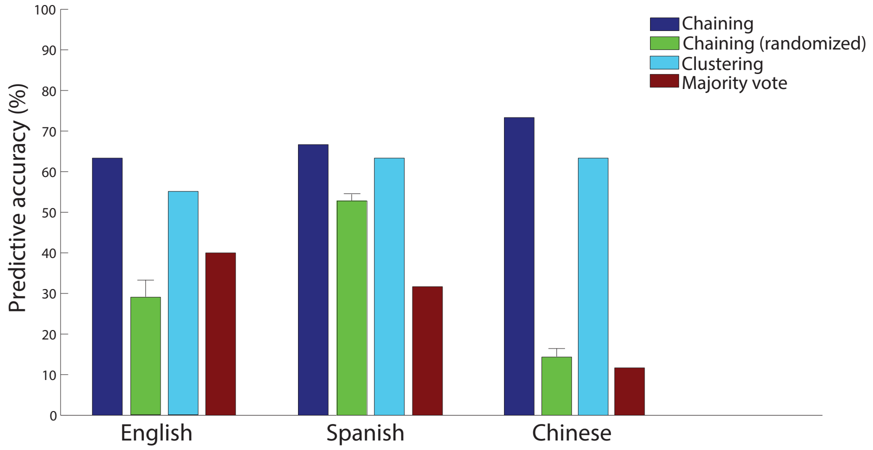

Figure 4: Summary of historical analysis of chaining. Error bars (standard error) are shown only for the case of randomized chaining, as that is the only case in which multiple simulations were run.

able, we used the model to predict its category membership, based only on items already encountered and holding out the category labels for the current item and all upcoming items. Whenever the model mis-predicted the category membership of an item due to encountering a novel category (one not yet represented by already-considered exemplars), we introduced that category and thus expanded the repertoire of categories that the model had to choose from among beyond that point in time. We then asked which model best predicted the data, when presented in historical order.

Figure 4 summarizes the predictive accuracies of these models. The results show that the chaining model accounts for the data better than the two alternative models, supporting the idea that historical chaining is involved in the formation of these lexical categories. We also wished to test whether the good fit of the chaining model is dependent on the historical time-stamps provided by our corpus searches. To that end, we re-ran the chaining model, but with the temporal sequence of exemplars randomized. We ran 100,000 such randomized sequences for each attested language. For all three languages, the mean prediction of the chaining model on the randomized historical sequences was significantly worse than with the real historical sequence ($p < 0.01$), suggesting that the success of this model does reflect the actual historical emergence of these items.[4]

These results demonstrate that the chaining model accounts well for the diachronic development of the extension of container categories. This outcome supports the proposal that historical semantic chaining is involved in the formation of container lexical categories.

Figure 5 illustrates the historical semantic chaining in the development of the English category *bottle*, on our analysis.

[4]A separate followup analysis showed that the chaining model does not outperform the other models on all datasets, suggesting that its performance on the Malt, Sloman, Gennari, Shi, and Wang (1999) dataset is attributable to a genuine match between the data and the assumptions of the model.

We observe that chaining in this real-world category is interestingly different from the idealized case we considered earlier in Figure 2. Instead of forming a long chain, "natural" semantic chaining in this instance takes the form of short chains grounded in hub exemplars (e.g. *iodine bottle* and *baby bottle*), which form local clusters within the category. These short chains might help to preserve local clustering properties of a lexical category, in contrast with the idealized linear chain that can extend indefinitely with no local constraint. But whether such natural chaining structures indeed support efficient communication is a question we will test in the next study.

Study 2: Chaining and efficient communication

We have seen that semantic chaining has the potential to yield inefficient communication, and that the container naming data of Malt et al. (1999) show evidence of semantic chaining over time. Left unaddressed is whether this natural semantic chaining in fact prevents efficient, informative communication.

The data of Malt et al. (1999) have not previously been analyzed with respect to whether they support efficient communication about containers. We conduct that analysis here, separately for each of the three languages (English, Spanish, and Chinese), using the computational formulation of efficient communication specified above. As before, we took the category system of each language to be determined by the modal head noun that was used to label it in the naming data, and we took similarities $sim(i, j)$ to be determined by the pile-sort data. We estimated the need probability $n(i)$ for each item i using frequencies of container names (modifier + head noun) from the Google Ngram American English corpus; specifically, we took frequencies at year 1999 which matches the year of publication of the original work by Malt et al. (1999). We then applied Equations 1 through 3 to obtain the communicative cost for the container naming system in each language.

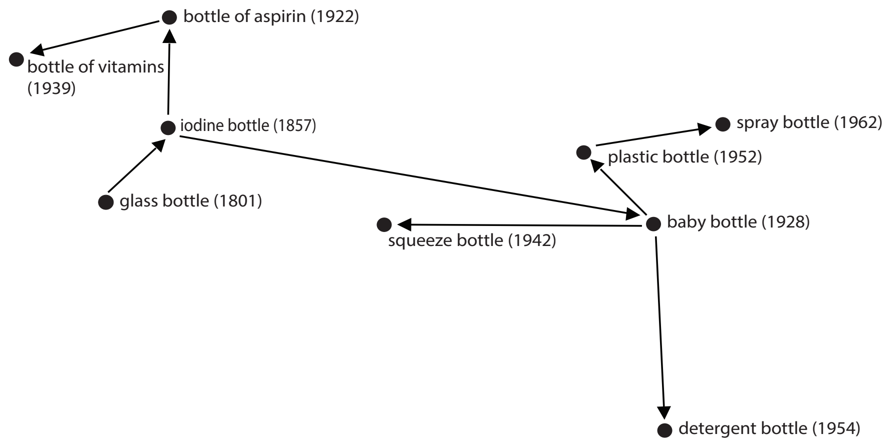

Figure 5: Semantic chaining in the English *bottle* category. Container names are annotated with dates identified from the corpus searches. Spatial proximity between two items roughly corresponds to their judged similarity. Arrows indicate the trajectory traced by the chaining model.

We take a category system to be near-optimally efficient if it is more informative than most hypothetical systems with the same number of categories. Thus, in order to assess the communicative efficiency of the English, Spanish, and Chinese container naming systems, we need to compare the communicative cost of each to the costs of a large set of hypothetical systems with the same number of categories as those reported by Malt et al. (1999) for this stimulus set (English: 7 categories; Spanish: 15 categories; Chinese: 5 categories). We also constrained the size of each category in a hypothetical system to be equivalent to that in the corresponding attested target system; this constraint ensures that attested and hypothetical categories are identical in number and size, and differ only in the exemplars that are assigned to those categories.

Concretely, for each target language (English, Spanish, Chinese), we constructed hypothetical comparison systems through simulated chaining in similarity space (cf. Khetarpal et al., 2013), as follows. We began by randomly choosing an initial exemplar and assigning it to an arbitrary category. We then extended that category to a new exemplar, which was selected by sampling exemplars in proportion to their similarity with the existing exemplar. We then repeated this chaining process, where the probability of a category being chosen for expansion was proportional to the number of remaining (as-yet-unlabeled) exemplars in that category. This process continued until we had assigned each exemplar to a category, such that each category had the same number of exemplars as in the attested target system. This procedure effectively generated a hypothetical chained system in the same similarity space as an attested system. Our use of chained systems as hypothetical competitors provides a conservative test, because it excludes from consideration unnatural-seeming hypothetical systems with disconnected (non-contiguous) categories.

For each of the three target languages, we created 100,000 such hypothetical chained systems, and we then compared the communicative cost of the attested target language to the costs of the hypothetical systems using the formulation presented earlier. The results are shown in Figure 6. Each of the three attested systems is significantly less costly than its corresponding class of hypothetical chained systems (English: $p < 0.001$; Spanish: $p < 0.0001$; Chinese: $p < .03$). We conclude from these results that although these systems do exhibit semantic chaining, each of them is nonetheless highly informative relative to a large class of comparable hypothetical chained systems.

Discussion

In this paper, we have presented two related contributions. First, we have provided what is, to our knowledge, the first computational demonstration of historical semantic chaining that is based in part on a large corpus of historical text. Second, we have shown that names for household containers in English, Spanish, and Chinese all support highly informative communication, despite the presence of historical chaining in this domain, and the potential of chaining to hamper informative communication. This result suggests that historical chaining may itself be constrained by the need for efficient, informative communication. Thus, the present work extends earlier analyses in the semantic domains of color, kinship, space, and number, not just by the addition of another domain, but also by the addition of a general phenomenon—historical chaining—that may be shaped by communicative need across domains.

Our findings leave a number of questions open. How general is the phenomenon of historical chaining, and the nature of it that we have suggested here? Does chaining appear in similar "hub exemplars plus short chains" form in other domains? Are there alternative models of chaining? How general is our finding that historical chaining may be constrained

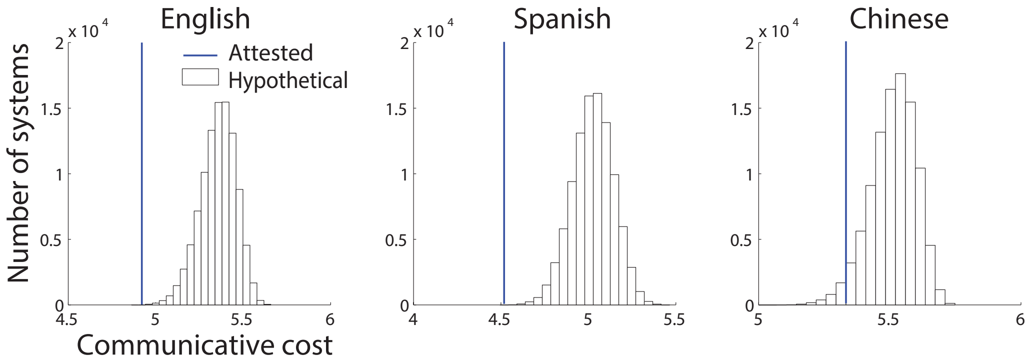

Figure 6: Efficiency analysis of container naming systems in English, Spanish, and Chinese. In each case, the attested system exhibits low communicative cost (high informativeness) relative to a large set of hypothetical systems of equal complexity.

by communicative forces? Future studies can address these questions by applying analyses similar to ours to other domains, as well as new analyses and computational models, to explore the linguistic packaging of meaning across languages and across time.

Acknowledgments

We thank Charles Kemp for earlier ideas on chaining and Rob Kass for the concept of change-point detection. Work on this project was funded by NSF award SBE-1041707 to the Spatial Intelligence and Learning Center (SILC).

References

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. University of California Press.

Brugman, C. M. (1988). *The story of over: Polysemy, semantics, and the structure of the lexicon*. Garland New York.

Bybee, J., Perkins, R., & Pagliuca, W. (1994). The evolution of grammar. *Tense, Aspect, and Modality in the Languages of the World*.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *PNAS*, *109*, 17897–17902.

Heit, E. (1992). Categorization using chains of examples. *Cognitive Psychology*, *24*(3), 341–380.

Hopper, P., & Traugott, E. (2003). *Grammaticalization*. Cambridge University Press.

Kass, R. E., Eden, U. T., & Brown, E. N. (2014). *Analysis of neural data*. Springer.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*, 1049–1054.

Khetarpal, N., Neveu, G., Majid, A., Michael, L., & Regier, T. (2013). Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *CogSci.*

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago.

Levinson, S. C., & Meira, S. (2003). 'Natural concepts' in the spatial topologial domain–adpositional meanings in crosslinguistic perspective: an exercise in semantic typology. *Language*, *79*, 485–516.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230–262.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., et al. (2011). Quantitative analysis of culture using millions of digitized books. *science*, *331*, 176–182.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *PNAS*, *108*, 3526–3529.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *PNAS*, *104*, 1436–1441.

Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Hoboken, NJ: Wiley-Blackwell.

Sloman, S. A., Malt, B. C., & Fridman, A. (2001). Categorization versus similarity: The case of container names. In U. Hahn & M. Ramscar (Eds.), *Similarity and categorization.* New York, NY: Oxford University Press.

Smith, K., Tamariz, M., & Kirby, S. (2013). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society.*

Xu, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the Cognitive Science Society.*