Language evolution in the lab tends toward informative communication

Alexandra Carstensen[1] (abc@berkeley.edu)
Jing Xu[4] (jing.xu@jhmi.edu)
Cameron T. Smith[2] (vmpfc1@berkeley.edu)
Terry Regier[2,3] (terry.regier@berkeley.edu)

Department of Psychology,[1] Department of Linguistics,[2] Cognitive Science Program[3]
University of California, Berkeley, CA 94720 USA

Department of Neurology, Johns Hopkins University, Baltimore, MD 21287 USA[4]

Abstract

Why do languages parcel human experience into categories in the ways they do? Languages vary widely in their category systems but not arbitrarily, and one possibility is that this constrained variation reflects universal communicative needs. Consistent with this idea, it has been shown that attested category systems tend to support highly informative communication. However it is not yet known what process *produces* these informative systems. Here we show that human simulation of cultural transmission in the lab produces systems of semantic categories that converge toward greater informativeness, in the domains of color and spatial relations. These findings suggest that larger-scale cultural transmission over historical time could have produced the diverse yet informative category systems found in the world's languages.

Keywords: Informative communication, language evolution, iterated learning, cultural transmission, spatial cognition, color naming, semantic universals.

The origins of semantic diversity

Languages vary widely in their fundamental units of meaning—the concepts and categories they encode in single words or other basic forms. For example, some languages have a single color term spanning green and blue (Berlin & Kay, 1969), and some have a spatial term that captures the notion of being in water (Levinson & Meira, 2003: 496), neither of which is captured by a single word in English. Yet at the same time, similar or identical meanings often appear in unrelated languages. What explains this pattern of wide yet constrained variation?

An existing proposal suggests an explanation in terms of the functional need for *efficient communication*: that is, communication that is highly informative yet requires only minimal cognitive resources. There may be many ways for systems to be communicatively efficient, and the different category systems that we see across languages may represent different language-specific solutions to this shared communicative challenge. This idea has accounted for cross-language semantic variation in the domains of color (Regier et al., 2007; 2015), kinship (Kemp & Regier, 2012), spatial relations (Khetarpal et al., 2013), and number (Xu & Regier, 2014).

However, this prior work has also left an important question unaddressed. In a commentary on Kemp and Regier's (2012) kinship study, Levinson (2012) pointed out that although that research explains cross-language semantic variation in communicative terms, it does not tell us "where our categories come from" (p. 989); that is, it does not establish what *process* gives rise to the diverse attested systems of informative categories. Levinson suggested that a possible answer to that question may lie in a line of experimental work that explores human simulation of cultural transmission in the laboratory, and "shows how categories get honed through iterated learning across simulated generations" (p. 989). We agree that prior work explaining cross-language semantic variation in terms of informative communication has not yet addressed this central question, and we address it here.

Iterated learning and category systems

The general idea behind iterated learning studies is that of a chain or sequence of learners. The first person in the chain produces some behavior; the next person in the chain observes that behavior, learns from it, and then produces behavior of her own; that learned behavior is then observed by the next person in the chain, who learns from it, and so on. This experimental paradigm is meant to capture in miniature the transmission and alteration of cultural information across generations; the learned behavior generally changes as it is filtered through the chain of learners.

Iterated learning and related learning studies have produced a number of findings that are directly relevant to the development of informative category systems. Kirby et al. (2008) showed that iterated learning of artificial languages resulted in those languages gradually becoming more structured, suggesting that linguistic structure could emerge from the dynamics of cultural transmission. Fedzechkina et al. (2012), in a non-iterated but relevant learning study, showed that learners of an artificial language restructured their input in a way that increases the efficiency of the learned system—specifically, learners preferentially deployed case marking in contexts in which it was highly informative, although that bias was not present in the input. This finding establishes the general principle that learners may alter their input in the direction of greater efficiency. However, the study did not examine the learning of systems

of semantic categories, and it is unknown whether the principle they established generalizes to the shaping of such systems. Finally, Xu et al. (2013) conducted an iterated learning study that *did* examine the learning of semantic category systems—but did not examine informativeness (see also Silvey et al., 2015). Xu et al. (2013) showed that iterated learning of color names produces systems of named color categories that are similar to those found in the world's languages. It is known that naturally-occurring color naming systems tend to support informative communication (e.g. Regier et al., 2015), so Xu et al.'s (2013) results indirectly suggest that iterated learning may lead to greater informativeness in category systems. However they did not directly test whether that is the case, and did not examine any semantic domain other than color.

Taken as a whole, the literature reviewed above leaves open two major relevant questions. (1) Does iterated learning of category systems in fact produce systems of greater informativeness? (2) If so, is this tendency toward informativeness found across different semantic domains? We pursue these questions here, to see whether they provide an answer to the challenge posed by Levinson (2012).

In what follows, we first present a computational framework for exploring semantic systems through the lens of informative communication. We then present two studies. In the first, we reanalyze the color naming data of Xu et al. (2013), and ask whether those data reveal convergence toward informative color naming systems. In the second study, we conduct an analogous iterated learning experiment in the domain of spatial relations, and ask the same question of those data. To preview our results, we find that in both domains, systems of semantic categories become increasingly informative through the process of iterated learning. We conclude that the informative yet varied systems of categories in the world's languages may have resulted from larger-scale processes of cultural transmission.

Informative communication

We take a semantic system to be *informative* to the extent that it supports accurate mental reconstruction by a listener of a speaker's intended message (Kemp & Regier, 2012; Regier et al., 2015). Figure 1 illustrates this idea in the context of communicating about color in English.

In the figure, time and causality flow from left to right. The speaker has in mind a particular target color *t* drawn from the universe *U* of all colors, shown here for simplicity as a 1-dimensional spectrum. The speaker represents this target color as a probability distribution *s* over *U*, centered at *t*. In our treatment below, we will assume that the speaker is certain of the target object, so that $s(t)=1$ and $s(i)=0\ \forall i \neq t$, but the framework can be generalized to accommodate speaker uncertainty about the target. The speaker wishes to communicate the target color to the listener, and so uses a word *w*: here, the English word *blue*. Having heard this word, the listener then attempts to mentally reconstruct the speaker's representation *s*, given *w*. The listener's reconstruction is also a probability distribution, *l*, and is intended to approximate the speaker's distribution *s* but is necessarily less precise, because the word *w* is semantically broad.

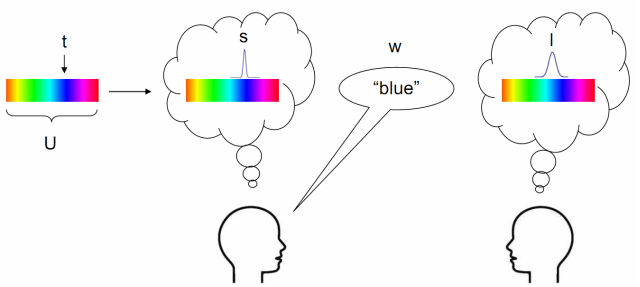

Figure 1: A scenario illustrating informative communication. From Regier et al. (2015).

The listener distribution is determined in different ways for different semantic domains, depending on the character of the domain. In the color and space analyses below, as in our earlier work in these domains (Regier et al., 2007; 2015; Khetarpal et al., 2013), we assume a similarity-based listener distribution: the listener reconstructs the speaker's intended meaning by assigning mass to each object *i* in the domain (here, each color *i*) as a function of how similar *i* is to the objects in the category named by *w*:

$$l(i) \propto \sum_{j \in cat(w)} sim(i, j) \qquad (1)$$

This captures the intuition that category-central referents (those with high similarities to other members) are the most expected targets when that category is used. The similarity $sim(i,j)$ between objects *i* and *j* is determined separately for different domains, as described in our studies below.

Given the speaker *s* and listener *l* distributions, we define the communicative cost $c(t)$ of communicating object *t* under a given semantic system to be the information lost in communication: that is, the information lost when *l* is taken as an approximation to *s*. We formalize this as the Kullback-Leibler divergence between *s* and *l*. In the case of speaker certainty as assumed here, this quantity reduces to surprisal:

$$c(t) = D_{KL}(s \,\|\, l) = \sum_{i \in U} s(i) \log_2 \frac{s(i)}{l(i)} = \log_2 \frac{1}{l(t)} \qquad (2)$$

Finally, we define the communicative cost for the domain as a whole to be the expected communicative cost over all objects in the domain universe *U*:

$$E[c] = \sum_{i \in U} n(i) c(i) \qquad (3)$$

Here $n(i)$ is the probability that the speaker will wish to talk about object *i*. In the analyses below, as in our earlier work in color and space (Regier et al., 2007; 2015; Khetarpal et al., 2013), we assume for simplicity that $n(i)$ is uniform. We take a semantic system to be informative to the extent that it exhibits low $E[c]$. A system could increase its informativeness through the addition of more categories; in our analyses we control for this possibility by comparing (groups of) systems with the same number of categories.

Study 1: Color

Xu et al. (2013) showed that iterated learning of color naming yields categorical partitions of color space that are similar to color naming systems found in the world's languages. They measured the distance between color categories produced in their experiment and those in the World Color Survey (WCS: Cook et al., 2005), the largest existing publicly available database of color naming data, containing color naming data from speakers of 110 languages of non-industrialized societies. Xu et al. (2013) found that as color naming systems in their iterated learning task were transmitted across generations of learners, the systems became more similar to those in WCS languages. In a separate study, Regier et al. (2015) assessed the communicative cost of color naming systems in the languages of the WCS, using the formal framework described above, and showed that the majority of these systems are highly informative, despite their diversity.

Taken together, these earlier findings suggest that color naming systems produced under iterated learning may come to resemble those found in languages through gradual increases in informativeness over generations. However, that proposal of increasing informativeness under iterated learning has not been directly tested. We test it here, by reanalyzing the color naming data from Xu et al. (2013)'s iterated learning experiment in terms of the framework described above.

Methods

Iterated learning of color. Xu et al. (2013) trained an initial generation of 20 participants on random partitions of color space into 3-6 categories, and then asked them to recall those categories by labeling a set of color chips accordingly. The next set of 20 participants each studied the assignment of labels to color chips of a single first generation learner, and created their own labelings in turn, which were then used to train the subsequent generation. This procedure was iterated over 20 chains of learners with 13 generations of learners each. In each generation of each chain, participants created a full color naming system by assigning a category label to each of the 330 color chips in the color naming array used in the WCS. Xu et al. then measured the dissimilarity between these transmitted category systems, at each generation, and the color naming systems of the WCS. They measured dissimilarity using variation of information (VI: Meilă, 2007), a distance measure between different groupings of the same set of items.

The data in Figure 2 (red line, left y-axis) are from Xu et al. (2013). These data show that as color naming systems are filtered through generations of learners, they become more similar to the natural systems of the WCS, as Xu et al. reported. We wish to ascertain whether this change also reflects a gradual increase in informativeness, brought about through transmission.

Communicative cost. In order to assess the informativeness of a given color naming system, we need to specify how similarity is determined in that domain (recall Equation 1). As in earlier work in this domain (Regier et al., 2007; 2015), we take the similarity of two colors i and j to be a Gaussian function of the perceptual distance between them:

$$sim(i, j) = \exp(-c \times dist(i, j)^2) \qquad (4)$$

Following Regier et al. (2007; 2015), the scaling factor c is set to .001 for all analyses reported here, and $dist(i,j)$ is the distance between colors i and j in the CIELAB color space. Given this, we can now assess the informativeness of a given color naming system following Equations 1-4.

Results

Figure 2 (blue line, right y-axis) shows the average communicative cost $E[c]$ of the 20 color naming systems in Xu et al's (2013) study, over the 13 generations of that study. Generation 0 corresponds to the random initial partitions supplied to the first generation of participants in training.

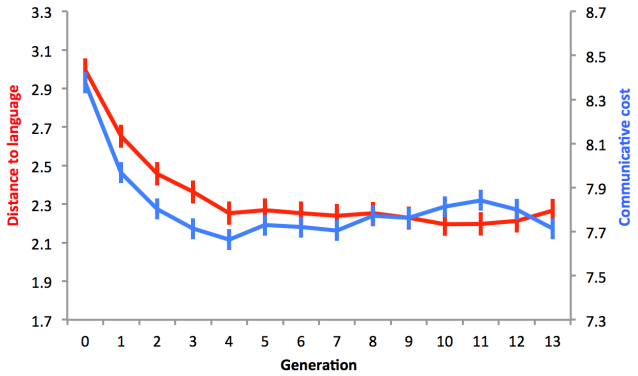

Figure 2: Average distance to WCS languages (red; left y-axis), and communicative cost (blue; right y-axis) of artificial systems of color categories, over generations of iterated learning. Bars indicate standard error of the mean.

It can be seen that these color naming systems exhibit decreasing communicative cost (increasing informativeness) over the first four generations of learners, after which no further systematic change is seen. This pattern of change over time closely parallels that seen in the similarity of lab-generated color naming systems to those of actual languages (red line). This finding suggests that artificial color naming systems come to resemble those found in languages through a transmission process that favors systems of greater informativeness.

Study 2: Spatial relations

Does iterated learning lead to increasing informativeness across multiple domains, or only in the domain of color? To

answer this question, we conducted an analogous study in a different semantic domain, that of spatial relations.

Languages categorize the spatial domain in a wide variety of ways that nonetheless show certain recurring tendencies (e.g. Levinson & Meira, 2003). Figure 3 gives a quick sense for this variation.

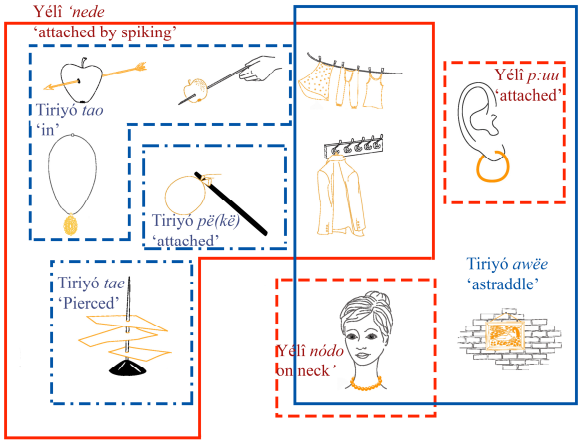

Figure 3: Ten spatial relations, as categorized in two languages: Tiriyó and Yélî-Dnye.
Adapted from Levinson & Meira (2003).

Additionally, spatial systems across languages tend to support informative communication (Khetarpal et al., 2013). In both of these respects, space is like color. However it is unlike color in that it is more complex. Perceptual color space is defined with respect to just three dimensions: hue, saturation, and lightness. In contrast, the mental representations underlying the kinds of spatial relations shown in Figure 3 appear to rely on a much wider range of spatial features (Levinson & Meira, 2003; Xu & Kemp, 2010).

We considered spatial naming data, collected both in the field and in the lab, relative to a standard stimulus set: the Topological Relations Picture Series (TRPS: Bowerman & Pederson, 1992). The spatial scenes in Figure 3 above are from the TRPS. The full TRPS is a set of 71 such line drawings depicting different spatial relations. Each image shows an orange figure object located relative to a black background object. We wished to discover whether iterated learning of category systems over these stimuli would converge toward the spatial systems of natural languages, and toward greater informativeness, in a parallel to the color findings reported above.

Methods

Iterated learning of spatial relations. 50 undergraduates at UC Berkeley took part in the study in return for class credit, forming 5 transmission chains of 10 generations each. Each participant completed an iterated learning task in which they studied and then attempted to recall category assignments for 4-category partitions of the 71 TRPS scenes.

Participants were instructed to learn spatial categories from an "alien language" by observing a series of scenes paired with visual sentences. In each training trial, a scene from the TRPS was presented for 5 seconds along with a visual sentence describing that scene in a hypothetical alien language. The visual sentence consisted of three smaller images beneath the main scene, as shown in Figure 4. The visual sentences showed the figure and ground objects from the main scene separately, and a colored patch indicating the alien spatial category to which the spatial relationship between figure and ground belongs. For example, in Figure 4, the participant is labeling the spatial relation apple-in-bowl as belonging to the category marked by red. Other scenes would be labeled by other colors, for a total of four color-coded categories.

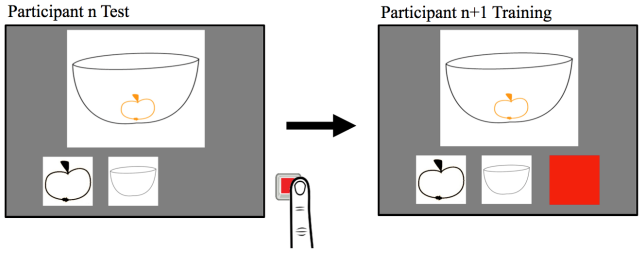

Figure 4: Example test and training trials from two consecutive generations of a transmission chain.

Participants completed two training sessions in which each of the 71 TRPS scenes was presented one at a time in random order paired with a color representing the spatial category to which that scene belongs. After two rounds of training, participants were shown the scenes and visual sentences a final time, but without the color label, and categorized each spatial relationship according to the alien language by pressing colored keys to indicate category assignments. Color labels and their locations on the keyboard were counterbalanced across participants within each iterated learning chain.

As in Xu et al.'s (2013) study, each of the 5 chains was initialized as a random partition of the 71 TRPS scenes into four roughly equally-sized categories, which the first participants in each chain studied during training and attempted to reproduce in the following test session. All subsequent participants in each chain were trained on the responses of the previous participant and were instructed to reproduce them as closely as possible, but were not aware that any of the data had any connection to other participants.

We excluded any participants whose categorization accuracy was at or below chance or who reported that they relied principally on non-spatial information (e.g. the objects involved) to learn the spatial categories.

Distance to languages. Analogous to Xu et al. (2013), we measured the dissimilarity between these transmitted spatial category systems at each generation, and the spatial systems of languages. Our target languages were a convenience sample: Arabic, Basque, Chichewa, Dutch, English, Japanese, Maijiki, Mandarin Chinese, and Spanish. All the

spatial naming data we drew on from these languages are unpublished. The data were collected either by our group (Arabic, Chichewa, Japanese, Mandarin Chinese, Spanish), or by collaborators who kindly shared their data with us and whom we gratefully acknowledge below (remaining languages). All data were collected relative to the TRPS scenes. For each language, we assigned to each TRPS scene the spatial term that was applied to that scene by the plurality of native speakers interviewed. This procedure yielded labels for all TRPS scenes, in each language. Following Xu et al. (2013), we used variation of information (VI) to measure the distance between category systems obtained through iterated learning, and those found in these languages.

Communicative cost. In order to assess informativeness for spatial relations, as for color, we needed an independent measure of similarity. We took the similarity between any two spatial relations stimuli to be determined by pile-sorting of those stimuli in a separate study. Khetarpal et al. (2010) asked native English speakers to sort the TRPS scenes into piles based on the similarity of the spatial relationships they depict. We took the similarity of any two scenes to be the proportion of participants who sorted those two scenes into the same pile in Khetarpal et al.'s (2010) data. Given this specification of similarity, we assessed the informativeness of spatial naming systems following Equations 1-3.

Results

Figure 5 (red line, left y-axis) shows the average distance (VI) between the spatial naming systems generated through iterated learning, and those of our language sample. This distance gradually decreases, as the systems are shaped by transmission from generation to generation. Thus, as in the case of color, iterated learning leads to spatial naming systems that become increasingly similar to those of natural languages.

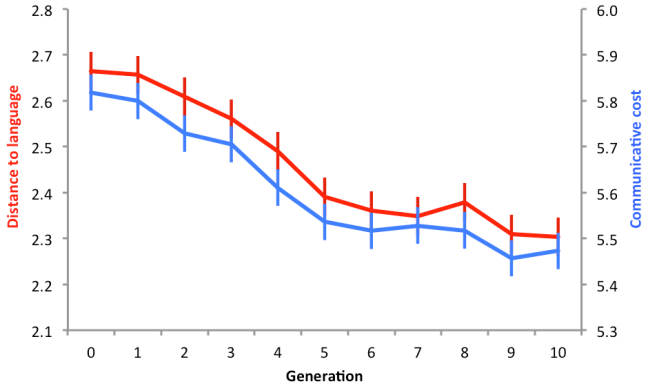

Figure 5: Average distance to languages (red; left y-axis), and communicative cost (blue; right y-axis) of artificial systems of spatial categories, over generations of iterated learning. Bars indicate standard error of the mean.

For comparison, Figure 5 (blue line, right y-axis) shows the average communicative cost of category systems across generations in our experiment. As in the case of color, this quantity also decreases as systems are transmitted from generation to generation, showing that transmitted spatial systems become more informative as they are transmitted. Moreover, again as in the case of color, this decrease closely tracks the decrease in distance to language, suggesting that iterated learning produces spatial systems that resemble those of languages through a transmission process that favors informative categories.

A natural concern is that the participants in our experiment may have been influenced by their knowledge of English, and that the increasing proximity of the learned systems to those of actual languages may have been driven by English semantic structuring. We feel this concern should be lessened by three observations (not shown in the figure): (1) the learned category systems get progressively closer to all languages considered, including those with categories that cross-cut English spatial terms; (2) the learned category systems are closer to some other languages (e.g. Arabic, Chichewa, and Mandarin Chinese) than they are to English; and (3) the same qualitative results obtain when English is excluded from the set of languages to which the learned category systems are compared. Given this, it seems plausible that the increasing proximity to languages may have been driven in large part by universal semantic tendencies and cognitive forces, rather than by the English language itself.

Increases in both informativeness and language-like semantic structuring are illustrated below in Figure 6. The figure shows scenes from a single category at the beginning (left panel) and end (right panel) of our experiment. After transmission through 10 generations of learners, the meaning of the category has been altered through the loss of many initial members depicting a wide variety of spatial relations, down to a set of scenes exemplifying a novel relational category that expresses the notion "tightly around", or encirclement and tight fit. This spatial notion is intuitively clear, yet does not correspond to a single spatial term in English, the primary language of our participants.

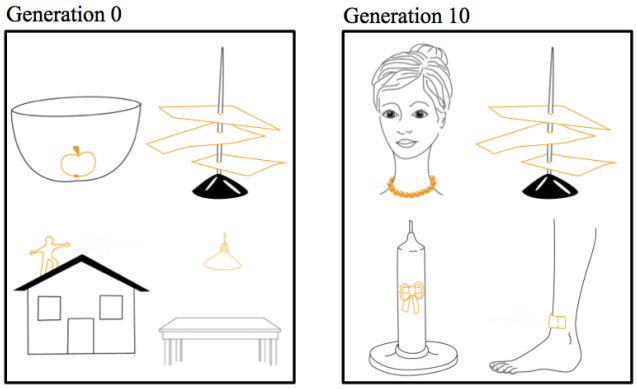

Figure 6: Representative scenes showing the semantic reorganization of a single category over transmission.

Discussion and conclusions

We have shown that iterated learning produces semantic systems that tend toward informative category structure, and also toward similarity with human languages. We find this pattern in two domains—color and spatial relations—suggesting that it may hold more generally across domains. To the extent that these findings do generalize, they suggest an answer to Levinson's (2012) question of how diverse category systems across languages assume their highly informative character.

A number of questions are left open by our findings. Would similar findings have been obtained if we had made other, but still reasonable, assumptions in our formalization of informative communication? Do these results extend to other semantic domains? Perhaps most importantly, do the results scale up to transmission in a larger social context? These questions are left open for future research. Despite these caveats, however, our initial findings reported here do suggest support for a specific account of the origins of the semantic diversity seen in the world's languages, as a natural result of shared communicative principles, operating across communities of language learners, and across time.

Acknowledgments

We thank Shubha Guha for assistance with early piloting and design of the experimental paradigm; Asifa Majid, Naveen Khetarpal, Grace Neveu, and Lev Michael for kindly sharing their spatial naming data; Vanessa Matalon, Ana Cuevas, Maggie Soun, Katie Chen, and Aaliyah Ichino for help in collecting additional spatial naming data; and Yang Xu and Joshua Abbott for valuable comments. We also thank Thomas Griffiths for help procuring data archives and Naveen Khetarpal for sharing pile-sort data and software. This work was supported by NSF under grant SBE-1041707, the Spatial Intelligence and Learning Center (SILC), and under NSF Graduate Research Fellowship grant DGE 1106400.

References

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

Bowerman, M., & Pederson, E. (1992). Cross-linguistic studies of spatial semantic organization. In *Annual Report of the Max Planck Institute for Psycholinguistics 1992* (pp. 53-56).

Cook, R., Kay, P., & Regier, T. (2005). The World Color Survey database: History and use. In H. Cohen and C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 223-242). Amsterdam: Elsevier.

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

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

Khetarpal, N., Majid, A., Malt, B., Sloman, S., & Regier, T. (2010). Similarity judgments reflect both language and cross-language tendencies: Evidence from two semantic domains. In S. Ohlsson and R. Catrambone (Eds.), *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

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, and I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS, 105*, 10681-10686.

Levinson, S.C. (2012). Kinship and human thought. *Science, 336*, 988-989.

Levinson, S.C. & Meira, S. (2003). 'Natural concepts' in the spatial topological domain—Adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language, 79*, 485-516.

Meilă, M. (2007). Comparing clusterings: An information based distance. *Journal of Multivariate Analysis, 98*, 873-895.

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: Hoboken, NJ: Wiley-Blackwell.

Silvey, C., Kirby, S. & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science, 39*, 212-226.

Xu, J., Dowman, M., & Griffiths, T. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B, 280*, 20123073.

Xu, Y. & Kemp, C. (2010). Constructing spatial concepts from universal primitives. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd 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 et al. (Eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.