Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses

Naveen Khetarpal (khetarpal@uchicago.edu)[a]
Grace Neveu (gkneveu@berkeley.edu)[b]
Asifa Majid (a.majid@let.ru.nl)[c]
Lev Michael (levmichael@berkeley.edu)[b]
Terry Regier (terry.regier@berkeley.edu)[b,d]

[a]Department of Psychology, University of Chicago, Chicago, IL 60637 USA
[b]Department of Linguistics, University of California, Berkeley, Berkeley, CA 94720 USA
[c]Center for Language Studies, Radboud University, Nijmegen, Netherlands
[d]Cognitive Science Program, University of California, Berkeley, Berkeley, CA 94720 USA

Abstract

Why do languages have the categories they do? It has been argued that spatial terms in the world's languages reflect categories that support highly informative communication, and that this accounts for the spatial categories found across languages. However, this proposal has been tested against only nine languages, and in a limited fashion. Here, we consider two new languages: Maijiki, an under-documented language of Peruvian Amazonia, and English. We analyze spatial data from these two new languages and the original nine, using thorough and theoretically targeted computational tests. The results support the hypothesis that spatial terms across dissimilar languages enable near-optimally informative communication, over an influential competing hypothesis.

Keywords: Spatial terms; semantic universals; informative communication; language and thought; semantic maps.

Spatial categories across languages

Spatial terms across languages often pick out different categories, as illustrated in Figure 1. Yet at the same time similar or comparable categories often recur across unrelated languages.

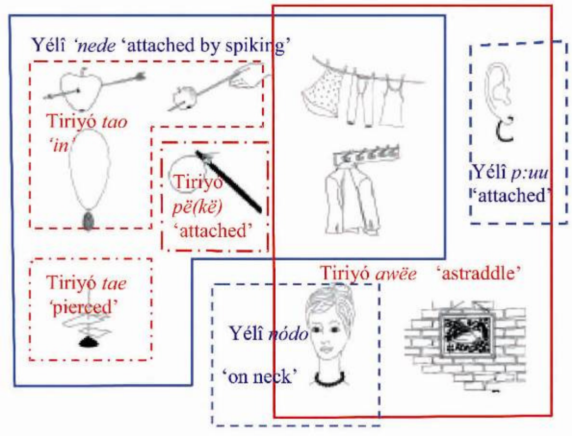

Figure 1: 10 spatial scenes, as categorized in 2 languages: Tiriyó and Yélî-Dnye. Source: Levinson et al. (2003).

A central question in cognitive science is why languages have the categories they do – in this case, why spatial categories exhibit the constrained cross-language variation they do (Bowerman & Pederson, 1992; Bowerman, 1996; Talmy, 2000; Levinson et al., 2003).

Informative communication

Recently, an answer to this question has been proposed that is grounded in general communicative principles. Khetarpal, Majid, & Regier (2009) argued that across languages, spatial categories are shaped by the need to support *informative communication*. On this view, the many different spatial systems observed across languages represent different means to this same end. This argument mirrors analogous arguments that have recently been advanced for the semantic domains of color (Regier, Kay, & Khetarpal, 2007) and kinship (Kemp & Regier, 2012), and also reflects a more general recent focus on informative communication as a central force that explains why languages take the forms they do (e.g. Fedzechkina, Jaeger, & Newport, 2012; Piantadosi, Tily, & Gibson, 2011).

Khetarpal et al. (2009) considered the 71 spatial scenes of the TOPOLOGICAL RELATIONS PICTURE SERIES or TRPS (Bowerman & Pederson, 1992), illustrated in part in Figure 1, as named by speakers of 9 unrelated languages: Basque, Dutch, Ewe, Lao, Lavukaleve, Tiriyó, Trumai, Yélî-Dnye, and Yukatek (Levinson et al., 2003). Each of these languages groups TRPS scenes together into language-specific spatial categories, and Khetarpal et al. (2009) asked whether these attested groupings support near-optimally informative communication. In a series of computational simulations, they asked whether each of these linguistic spatial systems supports informative communication better than a comparison class of hypothetical systems. They found that this is indeed the case. They concluded that spatial terms across languages reflect near-optimally informative spatial categories, and that this functional force may help to explain which spatial categories appear in the world's languages.

However, this earlier work is limited in three important respects. First, it considered data from only nine languages.

Such data are difficult and time-consuming to collect, and we are grateful to our colleagues at the MPI Nijmegen for sharing their data with us. Still, this is a very small sample, so it is possible that other languages would break the generalization made on the basis of these nine. Second, the earlier work tested the near-optimality claim against these nine languages in a narrow and limited way. Each language was compared to only 69 hypothetical systems that were intended to be comparable to it. Thus it is possible that many other, unexamined hypothetical systems may exist that are more informative than the attested system – again potentially breaking the generalization and undercutting the central theoretical claim. Third, the earlier work did not test the informativeness proposal against alternative explanations for constrained semantic variation.

Here we bring new data and analyses to bear on the claim that spatial categories across languages support informative communication, and that this force may account for the observed variation in spatial systems. The new data are from Maijɨki, an under-documented language of Peruvian Amazonia, and English. The new analyses compare eleven languages (Maijɨki, English, and the nine languages from Levinson et al., 2003) to much larger and more theoretically targeted sets of hypothetical systems. Critically, unlike the earlier analyses, the new analyses explicitly pit the claim of near-optimal informativeness against the competing and influential theoretical claim that semantic categories tend to pick out *connected regions* of conceptual or perceptual space (e.g. Croft, 2003; Haspelmath, 2003; Roberson, Davies, & Davidoff, 2000; Roberson, 2005).

In what follows we first describe Maijɨki and its spatial system, comparing it with that of English. We then lay out the hypotheses to be tested, our analyses of the eleven languages under consideration, and the results of these analyses. We conclude from these results that spatial systems across languages do indeed reflect near-optimally informative categories, and that this proposal is supported over the competing claim that categories pick out connected regions of conceptual or perceptual space. We suggest that the functional goal of informative communication may account for the wide but constrained variation found in spatial systems across languages.

Maijɨki

Maijɨki is an under-documented Western Tukanoan language of Peruvian Amazonia, spoken in the *departmento* of Loreto, near the Colombian-Peruvian border. The language is spoken by approximately 100 individuals, of whom some 25 are Maijɨki-dominant, although there are no monolingual speakers. The language is currently being documented as part of the Maijɨki Project, a multi-year effort to produce a grammar, text collection, and dictionary of the language (Michael, Beier, & Farmer, 2012). Maijɨki is unrelated to the other languages that we consider in this paper.

The spatial system of Maijɨki has only recently been investigated, and is described in detail by Neveu and Michael (in preparation). Spatial meanings are conveyed in Maijɨki by several means, including spatial adpositions and spatial verbs. For simplicity we focus on the major spatial adpositions, listed in Table 1 (tone marks are suppressed here and elsewhere in this paper).

Table 1: Spatial adpositions in Maijɨki.

Adposition	**Approximate meaning**
guibɨ	under
gunu	near an edge
ɨmɨjai	on top or above
jeteruru	behind
sanu	inside at bottom

The extensions of these Maijɨki spatial adpositions are illustrated in Figure 2 below, as subsets of the full set of 71 scenes in the TRPS. Also shown for comparison are spatial categories in English. In each of the 71 scenes, the figure object is shown in orange, the ground object in black, and the corresponding spatial meaning is the spatial relation between the figure and the ground. As can be seen, the spatial categories of Maijɨki differ from those of English. We seek general principles that help to determine which logically possible groupings of scenes constitute categories that are attested in the world's languages.

Hypotheses

We consider two hypotheses, which our analyses pit against each other, using data from Maijɨki, English, and the nine languages of Levinson et al. (2003).

Near-optimally informative communication

The first hypothesis is the one sketched above: that spatial categories across languages appear as they do because these categories maximize or near-maximize the informativeness of communication. We take a communicative system to be informative to the extent that it supports *accurate mental reconstruction* by a listener of a speaker's intended meaning (cf. communication accuracy: Lantz & Stefflre, 1964). This general idea, which also applies to other semantic domains, can be made concrete through the following communicative scenario.

A speaker has a particular spatial relation in mind, and wishes to communicate it to a listener. To that end, the speaker produces a spatial term that describes this spatial relation. The listener must then mentally reconstruct the original spatial relation that the speaker intended, from the term used. Because the listener knows only that the intended spatial relation falls in the general category named by the spatial term, the listener's mental reconstruction is the set of all spatial relations that are named by the term. We define the *reconstruction accuracy* to be the similarity of this mental reconstruction to the original intended spatial relation. In general, we hold that informative categories, and informative systems of categories, are those that support high reconstruction accuracy.

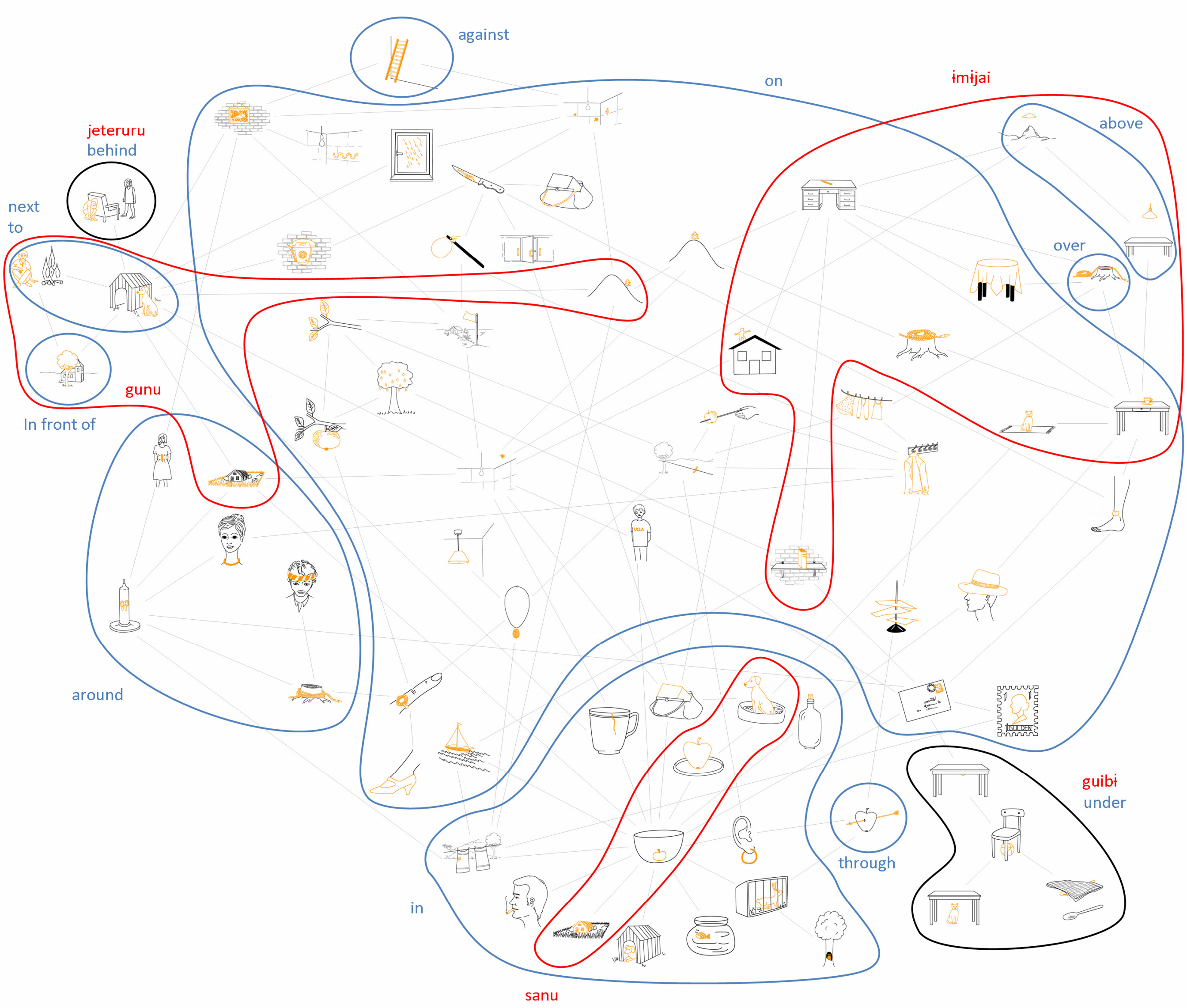

Figure 2: A semantic map showing spatial categories from Maijiki (red) and English (blue). Categories that appear in both languages are shown in black. Links connect scenes that are presumed to be universally related across languages. All displayed categories in both Maijɨki and English pick out connected regions of the map.

We formalize these ideas as follows.[1] Let *S* be the set of all possible spatial relations (here approximated by the spatial scenes of the TRPS, or the subset of those scenes that are assigned names by a given language). Let *sim(x,y)* be the similarity between two spatial relations *x* and *y* (here, similarity is gauged empirically as described below, and ranges from 0 = completely dissimilar to 1 = maximally similar). Let *s* be the specific spatial relation the speaker intends to convey, let *t* be the spatial term used to describe that spatial relation, and let *cat(t)* be the category or set of all spatial relations described by *t*, including *s*. Finally, let *era(s)* be the expected reconstruction accuracy of scene *s*, i.e. the similarity between the target spatial relation *s* and the listener's reconstruction of that spatial relation, based on the speaker's spatial term *t*. This is the average, over all spatial relations *r* in the same named category *cat(t)* as *s*, of the similarity between *r* and *s*:

$$era(s) = \frac{1}{|cat(t)|} \sum_{r \in cat(t)} sim(r,s) \qquad (1)$$

[1] Khetarpal et al. (2009) used a slightly different formalization of these ideas. We use this one because it maps cleanly onto the communicative scenario sketched above, in which a listener tries to understand a speaker's meaning. The results reported below remain qualitatively unchanged if the original formalization is used instead.

The overall expected accuracy of reconstruction, over all possible stimuli, is then given by:

$$R = \frac{1}{|S|}\sum_{s \in S} era(s) \qquad (2)$$

R is a measure of how well a given communicative system supports informative communication. The first hypothesis we consider is that attested linguistic spatial systems will tend to exhibit high *R*, compared with hypothetical systems.

The semantic map connectivity hypothesis

The second hypothesis we consider holds instead that attested categories pick out *connected regions* of a universal network of meanings called a *semantic map* (e.g. Croft, 2003; Haspelmath, 2003). Figure 2, in which we saw the spatial systems of Maijɨki and English, is an example of a semantic map. Here the meanings are spatial meanings, represented by the spatial scenes of the TRPS. These spatial meanings are assumed to be universally available, and the links in the network represent presumed universally available connections between closely related spatial meanings. As we have seen, different languages often group these meanings into categories differently, and these language-specific groupings are also displayed in the map. Thus a semantic map represents both presumed universal semantic structure and language-specific parcelings of that structure.

The core idea behind a semantic map is that across languages, semantic categories will always pick out *connected regions* of the network. In other words, a category should correspond to a group of meanings (here, scenes) that are connected in the sense that one may travel from any meaning in the category to any other by repeatedly traversing links in the network. The semantic map in Figure 2 was inferred automatically (Regier, Khetarpal, & Majid, in press) to accommodate, as connected regions, the spatial categories of the nine languages of Levinson et al. (2003). As can be seen, this network generalizes well to Maijɨki and English: all the displayed Maijɨki and English spatial categories also pick out connected regions of this map, although Maijɨki and English were not considered in its construction.[2] This fact suggests that the inferred universal structure of this semantic map, and the criterion of connectedness implicit in it, may in fact be an important constraint on semantic categories across languages. Similar ideas emphasizing the importance of connectedness as a determinant of what makes a good or natural category may also be found elsewhere (e.g. Levinson et al., 2003; Roberson, Davies, & Davidoff, 2000; Roberson, 2005).

Goal of our analyses

It has been previously suggested (e.g. Croft, 2003: 138; Cysouw, 2001: 609; Regier et al., in press) that connectedness in a semantic map may be too loose a constraint on category shape, in part because it allows elongated categories with no clear central region; thus, semantic categories in actuality may tend to be more compact and coherent than is suggested by this constraint alone. However it has not yet been determined whether informativeness provides a better account of cross-language variation in semantic systems. The analyses we present below seek to answer this open question, by deliberately pitting informativeness and connectedness against each other.

Analyses

We reasoned with the following predictions. The informativeness hypothesis predicts that attested linguistic spatial systems will support informative communication more effectively than almost all hypothetical systems – even if those hypothetical systems all pick out connected regions of a semantic map. The connectedness hypothesis in contrast does not make this prediction. Instead, on that hypothesis, it is connectedness rather than informativeness that plays a privileged role in determining which possible systems are actually attested – and so the informativeness of an attested linguistic spatial system should not tend to be any greater than the informativeness of other connected hypothetical systems.

For this reason, in our analyses we compared the informativeness of an actual linguistic spatial system with that of hypothetical variants, all of which correspond to connected regions of the semantic map of Figure 2. If informativeness is a major determinant of attested category systems, we expect the actual linguistic spatial system to support informative communication better than the connected hypothetical variants.

Crawling a semantic map

We generated hypothetical connected variants of existing systems by randomly "crawling" a semantic map, by analogy with web-crawling – that is, through random graph traversal of a semantic map. We began with the semantic map in Figure 2, but with no labels assigned to the scenes. Then, for a given target language (e.g. English), we construct a hypothetical connected variant of that language as follows. Start by randomly selecting one the spatial terms in the language—call this term *t* and the number of scenes associated with it *k*. Now randomly select one of the scenes in the graph and label it *t*. Then select another scene at random from the set of as-yet-unlabeled scenes directly connected to some scene already labeled *t*, and label that new scene *t* as well; if there are no such scenes from which to select, the procedure terminates and begins again with no labels on any nodes. This step of extending the label *t* to neighboring scenes is repeated until there are *k* scenes

[2] Regier et al. (in press) presented slightly different extensions of English categories against this map, one of which was not connected. We have chosen these extensions instead because (1) they allow English categories to be connected in this map, (2) that connectedness allows us to include English in our upcoming analyses, and (3) these extensions agree well with our linguistic intuitions.

associated with *t*. The process as a whole is repeated for all terms in the language.

Methods

We conducted semantic-map-crawling analyses separately for each of the eleven languages under consideration: Maijɨki, English, Basque, Dutch, Ewe, Lao, Lavukaleve, Tiriyó, Trumai, Yélî-Dnye, and Yukatek. For each language, 2000 hypothetical connected variants were generated as described above, each with the same number of categories, and the same number of scenes per category, as the original. For each real or hypothetical spatial naming system, we calculated *R*, our measure of reconstruction accuracy, using equations 1 and 2 above. The categories *cat(t)* used to label specific scenes were determined by the naming system under consideration. The similarity of each pair of scenes *x* and *y*, *sim(x,y),* was determined empirically by pile-sorting. Khetarpal et al. (2009) had asked speakers of English and Dutch to sort the TRPS scenes into piles on the basis of the similarity of the spatial relation portrayed, and they took the similarity of any two scenes to be the proportion of all their participants who sorted those two scenes into the same pile.[3] We used the pile-sort-derived similarity judgments from that earlier study. For each language, we then compared the reconstruction accuracy *R* for the language itself to the distribution of *R* obtained for hypothetical connected variants of that system.

Results

Figure 3 below presents the results of our analysis of Maijɨki. The red line shows the informativeness (*R*) of the Maijɨki spatial adpositional system, and the blue histogram shows the frequency with which various values of *R* were exhibited by hypothetical connected variants of Maijɨki, obtained by randomly crawling the semantic map of Figure 2.

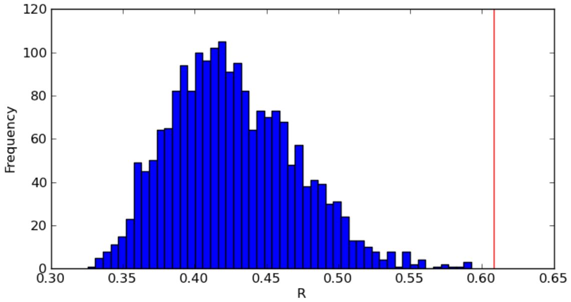

Figure 3: Informativeness of communication supported by the Maijɨki spatial adpositional system (red line), compared with that supported by 2000 hypothetical variants derived by randomly crawling a semantic map (blue histogram).

The actual Maijɨki system supports informative communication more effectively than any of the sampled hypothetical connected variants. These results are consistent with the claim that languages tend to have highly informative spatial systems, and that informativeness is more relevant to the shape of such systems than is connectedness. Similar results from other languages would strengthen this conclusion.

Figure 4 below presents analogous results for English. Again, the actual English system supports informative communication more effectively than any of the sampled hypothetical connected variants.

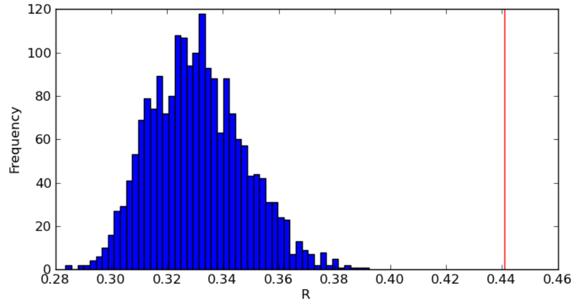

Figure 4: Informativeness of communication supported by the English spatial system (red line), compared with that supported by 2000 hypothetical variants derived by randomly crawling a semantic map (blue histogram).

Finally, Table 2 below presents summary results of semantic map crawling analyses for all eleven languages we consider. In this case, the results are given numerically, as the proportion of hypothetical variants that the actual linguistic system scores higher than in *R* (reconstruction accuracy). The results shown here for Maijɨki and English summarize the results from the histograms displayed above; for the remaining nine languages, we present results in summary form only, to conserve space. In all cases, the actual linguistic system outperforms most of the sampled hypothetical connected variants, and in several cases it outperforms all of them.

Table 2: Summary results of semantic map crawling analyses for all languages considered in this study.

Language	**Result**
Basque	> 99.95%
Dutch	> 100.00%
English	> 100.00%
Ewe	> 99.95%
Lao	> 96.20%
Lavukaleve	> 99.75%
Maijɨki	> 100.00%
Tiriyó	> 100.00%
Trumai	> 100.00%
Yélî-Dnye	> 97.35%
Yukatek	> 99.95%

In sum, each of the 11 languages considered supports informative communication more effectively than most sampled hypothetical variants of those systems – even when

[3] A followup study found that these pile-sorts were broadly similar across the two languages, although they did reflect the sorter's native language to some extent (Khetarpal et al., 2010).

the variants are connected regions of a semantic map. These results are consistent with the hypothesis that informativeness shapes category systems across languages, and that it does so more than connectedness in a semantic map.

Conclusions

Our findings support the claim that spatial systems across languages reflect the need for informative communication. They do so based on new evidence, including evidence from an under-documented language, and on new large-scale analyses that directly pit informativeness against the competing claim that natural categories pick out connected regions of a semantic map.

These findings also leave a number of issues unresolved, suggesting directions for future investigation. Theoretically, our analyses have focused on the informativeness of a given system, by comparing that system to competitors of comparable complexity – thus deliberately controlling for, and not investigating, the complexity of these systems. A more complete account would investigate both informativeness and complexity, and the tradeoff between these two general forces (e.g. Kemp & Regier, 2012). Empirically, eleven languages is still a small sample when considered relative to all existing languages. We feel that every new language considered adds important evidence, particularly under-documented languages such as Maijɨki – but consideration of more languages will allow more definitive conclusions.

Nonetheless, the present results lend substantial new support to the hypothesis that informativeness plays an important role in shaping spatial semantic systems across languages. In so doing, these results add to the current literature that suggests that the need for informative communication may be a key functional force that explains why languages have the forms that they do.

Acknowledgments

This work was supported by NSF under grants SBE-0541957 and SBE-1041707, the Spatial Intelligence and Learning Center (SILC). We thank Alex Carstensen for her help with the Maijɨki data.

References

Bowerman, M. (1996). Learning how to structure space for language: A cross-linguistic perspective. In P. Bloom, M. Peterson, M. Garrett, & L. Nadel (Eds.) *Language and space* (pp. 385–436). Cambridge, MA: MIT 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).

Croft, W. (2003). *Typology and universals: Second edition.* Cambridge, UK: Cambridge University Press.

Cysouw, M. (2001). Review of Martin Haspelmath, *Indefinite Pronouns* (1997). *Journal of Linguistics 37*, 607-612.

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

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: L. Erlbaum Associates

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In M. Tomasello (Ed.), *The new psychology of language, vol. 2* (pp. 211-242). Mahwah, NJ: Erlbaum.

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., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. In N. Taatgen et al. (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Lantz, D., & Stefflre, V. (1964). Language and cognition revisited. *Journal of Abnormal and Social Psychology, 69*, 472-481.

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

Michael, L., Beier C., & Farmer, S. (2012). Diccionario Bilingüe Maijɨki-Castellano y Castellano-Maijɨki. Maijɨki Project.

Neveu, G., & Michael, L. (in preparation). The semantics and pragmatics of topological spatial relations in Maijɨki. Ms.

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., Khetarpal, N., & Majid, A. (in press). Inferring semantic maps. *Linguistic Typology*.

Roberson, D., Davies I. & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129, 369-398.

Roberson, D., Davidoff, J., Davies, I.R.L., & Shapiro, L.R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology, 50*, 378-411.

Talmy, L. (2000). How language structures space. In L. Talmy (Ed.) *Toward a cognitive semantics, Volume 1* (pp. 177-254). Cambridge, MA: MIT Press.