Similarity judgments reflect both language and cross-language tendencies: Evidence from two semantic domains

Naveen Khetarpal (khetarpal@uchicago.edu)[a]
Asifa Majid (asifa.majid@mpi.nl)[b]
Barbara Malt (barbara.malt@lehigh.edu)[c]
Steven Sloman (steven_sloman@brown.edu)[d]
Terry Regier (terry.regier@berkeley.edu)[e]

[a]Department of Psychology, University of Chicago, Chicago, IL 60637 USA
[b]Max-Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands
[c]Department of Psychology, Lehigh University, Bethlehem, PA 18015 USA
[d]Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912 USA
[e]Department of Linguistics, Cognitive Science Program, University of California, Berkeley, CA 94720 USA

Abstract

Many theories hold that semantic variation in the world's languages can be explained in terms of a universal conceptual space that is partitioned differently by different languages. Recent work has supported this view in the semantic domain of containers (Malt et al., 1999), and assumed it in the domain of spatial relations (Khetarpal et al., 2009), based in both cases on similarity judgments derived from pile-sorting of stimuli. Here, we reanalyze data from these two studies and find a more complex picture than these earlier studies suggested. In both cases we find that sorting is similar across speakers of different languages (in line with the earlier studies), but nonetheless reflects the sorter's native language (in contrast with the earlier studies). We conclude that there are cross-culturally shared conceptual tendencies that can be revealed by pile-sorting, but that these tendencies may be modulated to some extent by language. We discuss the implications of these findings for accounts of semantic variation.

Keywords: Language and thought; semantic universals; linguistic relativity.

A universal basis for semantic variation?

The semantic systems of the world's languages vary considerably. This observation has suggested two opposed accounts of the relation between language and thought. The Sapir-Whorf hypothesis holds that such cross-language differences cause corresponding differences in cognition, leading speakers of different languages to think about and perceive the world substantially differently (Lucy, 1992; Majid et al., 2004; Roberson et al., 2000). In contrast, many other theories accommodate such variation by positing a universal conceptual space that is partitioned in different ways by different languages (Berlin & Kay, 1969; Croft, 2003:139; Levinson & Meira, 2003; Majid et al., 2008; Malt et al., 1999; Regier et al., 2007). On this view, the significant point about the variation is that many logically possible semantic configurations are never attested – thus, the constrained variation illuminates underlying commonalities in human cognition.

Although the starting point for this debate is linguistic – namely the observation of semantic diversity across languages – a natural means of testing it is by probing non-linguistic cognition. The Whorfian view predicts that speakers of languages with different semantic systems should conceive of the world differently, each group in line with their own language's semantic system. The universal-space view in contrast predicts that speakers of different languages should conceive of the world similarly.

One source of support for the universal-space view comes from *pile-sorting*. In the first large-scale quantitative study of its kind, Malt et al. (1999) asked speakers of English, Chinese, and Spanish to name a set of household containers – e.g. a jar, a juice-box, an ice-cream carton, etc. – and to pile-sort pictures of these items on the basis of their overall similarity. They found that while naming patterns differed substantially across languages, sorting patterns did not.

The same view is indirectly supported by recent studies that explain differing patterns of semantic structure in the world's languages as optimal or near-optimal partitions of an underlying and presumably universal similarity space. Regier et al. (2007) demonstrated that color naming in the world's languages is consistent with this idea, assuming a standard perceptual color space, CIELAB. This account explains universal tendencies in color naming while also accommodating some deviation from those tendencies, as is observed empirically. Khetarpal et al. (2009) showed that the same idea can account for semantic variation in the spatial domain. In the spatial case, however, no standard independent assessment of a universal similarity space exists. Therefore, inspired by the Malt et al. (1999) results, Khetarpal et al. (2009) based their analysis on similarities derived from pile-sorting of spatial scenes by speakers of Dutch and English. Critically, while they assumed that these similarities would be universal or near-universal, and while their results were consistent with that assumption, they did not directly test the assumption. We test it here.

To preview our results, we find that pile-sorting of spatial stimuli, according to the data of Khetarpal et al. (2009), is broadly similar across languages – but does nonetheless

differ as a function of language. These results were obtained using an analysis different from that of Malt et al. (1999) – thus the question arises whether Malt et al.'s (1999) container data would yield similarly mixed results under our analysis. We show that they do. We conclude that on one analysis at least, pile-sorting reveals not just shared cross-language tendencies, but also apparent influence of the sorter's native language, suggesting an interesting combination of the universalist and Whorfian positions (Regier & Kay, 2009).

Spatial language and cognition

Khetarpal et al. (2009) demonstrated a commonality underlying the diversity of spatial naming in the world's languages. They based their study on a set of 71 spatial scenes that were originally designed by Melissa Bowerman and Eric Pederson. Figure 1 shows a sample of 10 of these scenes, as categorized in 2 languages.

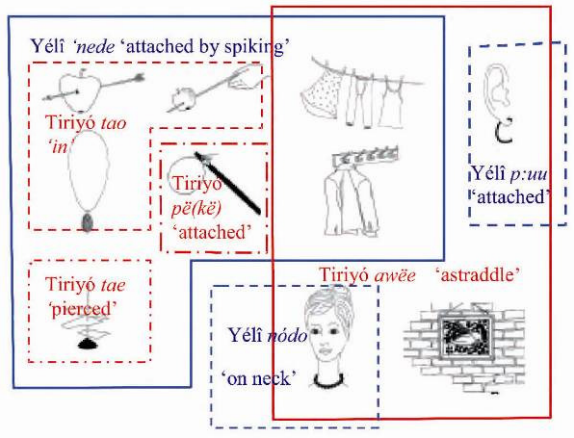

Figure 1: 10 spatial scenes, as categorized in 2 languages: Tiriyó and Yélî-Dnye. Source: Levinson & Meira (2003).

Khetarpal et al. (2009) had native speakers of Dutch and native speakers of American English sort pictures of these 71 spatial scenes into piles on the basis of the similarity of the spatial relation portrayed. Afterwards, they also elicited names for these spatial relations from each sorter in his or her native language. They then derived similarity judgments from sorting behavior: the similarity between any two scenes *x* and *y* was taken to be the proportion of all participants (American and Dutch pooled together) who sorted *x* and *y* into the same pile. Finally, they assessed the spatial semantic systems of 9 unrelated languages (one language was Dutch but the rest were unrelated to Dutch and English; Levinson & Meira, 2003) relative to these similarities. They found that these 9 attested spatial semantic systems maximized similarity within categories, and minimized it across categories (Garner, 1974), more than did a reasonable set of competitor systems of comparable complexity; in this sense these attested spatial semantic systems are *near-optimal*. This finding is consistent with the assumption that the sorting-derived similarities are universal – since they help to explain the spatial semantic systems of unrelated languages. But is this assumption in fact correct – or do these similarities reflect the sorters' native language? A natural means of testing this question is to compare the sorts produced by speakers of English and Dutch to the naming systems of the same two languages.[1] The Whorfian prediction is that speakers of each language should sort in a manner that reflects their native language, more than the other language. The universalist prediction is that speakers of the two languages should sort identically.

Methods

Naming data. For both English and Dutch, separately, we recorded the modal spatial term for each of the 71 spatial scenes — i.e. the spatial term that was used by the largest number of speakers of the language to name that scene. Ties were broken by random choice. The resulting labeling of the 71 scenes was taken to be that language's spatial naming system.

Sorting data. We analyzed the English and Dutch sorting data in 3 ways. First, we measured the *correlation* of sorting behavior across languages. Second, we measured how well sorts matched the semantic systems of English and Dutch, using *edit distance*. Third, we examined the *height*, or coarse-grainedness, of the sorts and of the English and Dutch semantic systems, since this quantity is helpful in interpreting other analyses, as will be seen below. Here, we describe each analysis in turn.

Correlation analysis. Following Malt et al. (1999), we compared sorts produced by English and Dutch speakers as follows. For each of Dutch and English, for each pair of scenes, we counted the number of times those two scenes were placed in the same pile by speakers of that language. This yielded, for each of the two languages, a vector of (71×70)/2 = 2485 co-sorting counts. We determined the correlation of the Dutch vector with the English vector.

Edit-distance analysis. We took a pile-sort of the 71 scenes to be a *partition* of those stimuli into groups; we similarly took a language's names applied to those scenes to be a partition of the same set of stimuli into groups. We quantified the dissimilarity between two such partitions by measuring the *edit distance* between them. The edit distance between two partitions A and B is the minimum number of operations required to change A into B, where each operation involves moving a single item from one group to another (possibly empty) group. We computed edit distances via the Hungarian algorithm for bipartite graph

[1] We collected new English data analogous to that of Khetarpal et al. (2009), since their English naming data were incomplete. We report here the comparison of Khetarpal et al.'s (2009) complete Dutch data with our complete English data. Comparison of Khetarpal et al.'s (2009) Dutch and English data yield qualitatively the same results as those we report here.

matching (Deibel et al., 2005).[2] For each pile sort produced by a speaker of either Dutch or English, we determined its edit distance to the partition defined by the Dutch language, and its edit distance to the partition defined by the English language.

Height analysis. The *height* of a partition is a measure of how coarse-grained it is: greater height indicates coarser grain, while lower height indicates finer grain. Height is defined as the sum, over all groups in a partition, of the number of pairs of items in each group (Coxon, 1999):

$$height = \sum_i \binom{g_i}{2} = \sum_i g_i(g_i - 1)/2$$

where g_i is the number of items in group i. We measured the height of the partitions corresponding to the English and Dutch naming systems, and the height of each pile-sort.

Results and discussion

Correlation. The correlation of the Dutch and English co-sorting vectors was 0.87. This correlation is fairly high, and is greater than the agreement between halves of the same group (Dutch or English): the mean within-group split-half reliability was 0.80. This result suggests that speakers of the two languages sorted quite similarly.

Edit distance. Edit distance gives us a means of measuring the dissimilarity between pile-sorts and naming systems. Figure 2 shows the average edit distance of sorts produced by Dutch speakers and those produced by English speakers, to the Dutch and English naming systems.

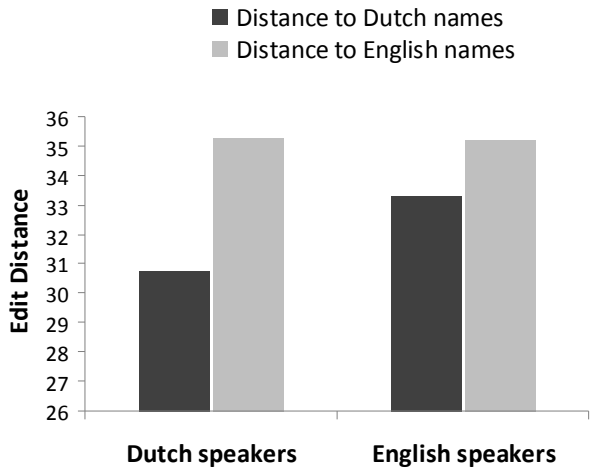

Figure 2: Edit distance of sorts, produced by Dutch and English speakers, to the Dutch and English naming systems.

We analyzed these data as follows. For each sorter from each of the two languages, we created a difference score: the edit distance of that person's pile sort to the English naming system minus the edit distance of that person's pile sort to the Dutch naming system. The difference scores for both groups were significantly greater than 0 (Dutch: M=4.5, $t(23) = 4.83$, $p < .0002$; English: M=1.92, $t(23) = 3.81$, $p < .002$), indicating that speakers of both languages sorted more in line with Dutch than with English. The Dutch mean difference score was greater than the English one ($t(46) = 2.44$, $p < 0.05$; all p values Bonferroni-corrected), indicating that Dutch speakers showed this preference for Dutch over English more strongly than English speakers did. Thus there appears to be both a cross-language tendency to sort more in line with Dutch than with English (a universalist finding), and a tendency to sort in line with one's native language (a Whorfian finding); these two forces pull in the same direction for Dutch speakers, but in opposite directions for English speakers.

What is it about the Dutch naming system such that speakers of both languages sort more in line with it than with English? It may be relevant that Dutch appears to be semantically *finer-grained* than English in this domain. For example, the English spatial term *on* covers a broad range of spatial meanings, including a cup on a table, and a picture on a wall – whereas these two spatial configurations are named differently in Dutch (as *op* vs. *aan*, respectively). Thus a possible explanation for the privileged status of Dutch in our results above is that people may tend to sort in a manner that is finer-grained than either language, and therefore more like the finer-grained language – in this case Dutch.

Figure 3 shows that this is the case. The height quantity measures the coarseness of a partition; thus, comparison of the two vertical lines shows that Dutch naming is indeed finer-grained than English naming with respect to these spatial scenes. Moreover, the bulk of sorts produced by speakers of both languages is finer-grained than the finer-grained language, Dutch.

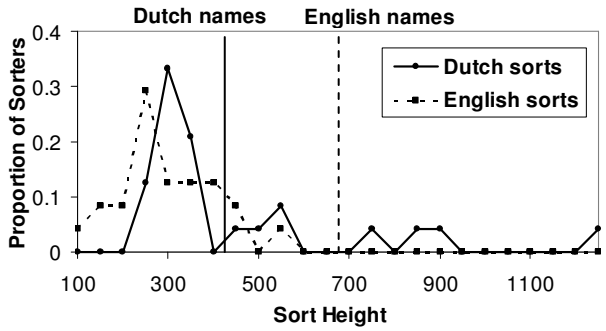

Figure 3: The height (coarse-grainedness) of the Dutch and English naming systems, and sorts produced by speakers of these two languages.

Thus, it seems likely that Dutch emerges as privileged in our edit-distance results at least in part because it is finer-grained than English in this domain. But are these results attributable to fine grain *per se*, or to the particular fine-grained partition that Dutch represents? To test this, we

[2] See http://psych.uchicago.edu/~khetarpal/code/edit-distance for our code, which extends an implementation written by Gary Baker and released under GPLv3.

also compared the pile-sorts to Dutch-*like* partitions which are as fine-grained as Dutch but group the items differently. The set of Dutch-like partitions was sampled repeatedly (n=3.5x10^6) by randomly grouping items such that the total number of groups equaled the number of Dutch spatial terms and the sizes of these groups matched the number of items associated with the Dutch spatial terms. We then measured the average edit distance from English speakers' sorts to each of these sampled hypothetical Dutch-like partitions (*Min*=46.79, *Mean*=52.09, *Max*=55.13), and the average edit distance from Dutch speakers' sorts to each of these sampled hypothetical Dutch-like partitions (*Min*=46.04, *Mean*=51.48, *Max*=54.29). In both cases the average edit distance of the sorts to actual Dutch (shown in Figure 2) was less than to any of the sampled hypothetical Dutch-like partitions of equally fine grain.[3] This finding suggests that the privileged status of Dutch in our edit-distance results is a function not just of its fine grain, but also of the similarity relations it captures.

Taken together, these reanalyses of the Khetarpal et al. (2009) spatial data suggest that spatial similarity judgments as gauged by pile-sorting are quite similar and fine-grained across languages – a universalist finding – but that they nonetheless vary in line with the sorter's native language – a Whorfian finding.

Container names and cognition

Our present analysis of the Khetarpal et al. (2009) spatial data revealed a mixed picture, in contrast with the purely universalist results of Malt et al. (1999) on containers. But our result was obtained through an edit-distance analysis that Malt et al. (1999) did not use. This raises the question whether the Malt et al. (1999) data would also exhibit an effect of language if analyzed using edit distance. We sought to test this question.

Malt et al. (1999) based their study on 60 pictures of simple containers, such as cartons, boxes, bottles, and the like. They asked speakers of 3 different languages – American English, Mandarin Chinese, and Argentinean Spanish – to name the containers shown in these pictures and to sort them into piles, on several different bases. Here, we re-examine their data from English and Chinese, for which data were readily retrievable, and we focus on pile-sorting based on overall similarity of the containers, rather than functional or perceptual similarity, which Malt et al. (1999) also probed. Importantly, while the semantic categories for the various containers differed across languages, the overall sorts showed no effect of language in their analyses.

Methods

We analyzed Malt et al.'s (1999) container naming and sorting data from Chinese and English using the same methods we had applied to the spatial data of Khetarpal et al. (2009). Specifically, we (1) identified each language's semantic partitioning of the space by determining the modal term applied to each stimulus in each language, and conducted (2) correlation, (3) edit-distance, and (4) height analyses of the sorting and naming data.

Results and discussion

Correlation. The correlation of the Chinese and English co-sorting vectors was 0.91, as Malt et al. (1999) had found. This correlation is quite high, and is comparable to the agreement between halves of the same group (Chinese or English): the mean within-group split-half reliability was 0.90. This result suggests that speakers of the two languages sorted quite similarly.

Edit distance. Figure 4 shows the average edit distance of sorts produced by Chinese speakers and those produced by English speakers, to the Chinese and English naming systems.

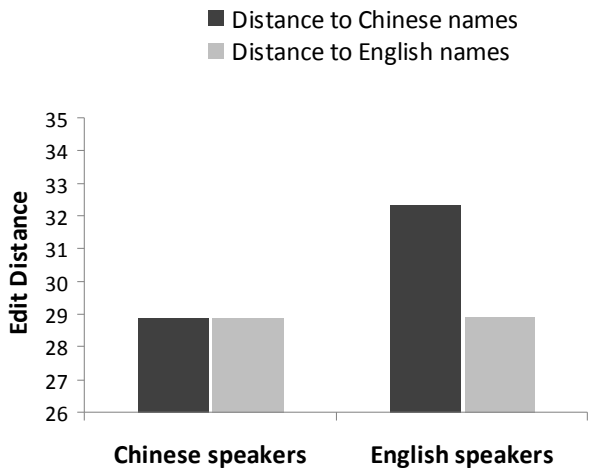

Figure 4: Edit distance of sorts, produced by Chinese and English speakers, to the Chinese and English naming systems.

We analyzed these data as before. For each sorter from each of the two languages, we created a difference score: the edit distance of that person's pile sort to the Chinese naming system minus the edit distance of that person's pile sort to the English naming system. The mean difference score for Chinese speakers was 0.0 (*SD* = 5.99), indicating that Chinese speakers sorted in a manner equally similar to the Chinese and English naming systems. In contrast, the mean difference score for English speakers was significantly greater than 0 (*M*=3.43; $t(55) = 6.17$, $p < .0002$), indicating that English speakers sorted in a manner more like the English than like the Chinese naming system. The English mean difference score was greater than the Chinese one ($t(36.6^4) = 2.64$; $p < .05$; all p values Bonferroni-corrected), indicating that English speakers sorted in line with English

[3] The actual Dutch naming system is also by definition a Dutch-like partition.

[4] Heteroscedasticity corrected using Welch's method.

more than Chinese to a greater extent than Chinese speakers did. As in the spatial case, a natural interpretation of these data is that there is a cross-language tendency to sort more in line with English than with Chinese, and also a tendency to sort in line with one's native language. For Chinese speakers these two forces cancel each other out, whereas for English speakers they reinforce each other.

Given our earlier discussion, a general tendency to sort more in line with English than with Chinese naming would make sense if English were more fine-grained than Chinese in this domain, and if people sorted more finely than either language. Figure 5 shows that this is the case.

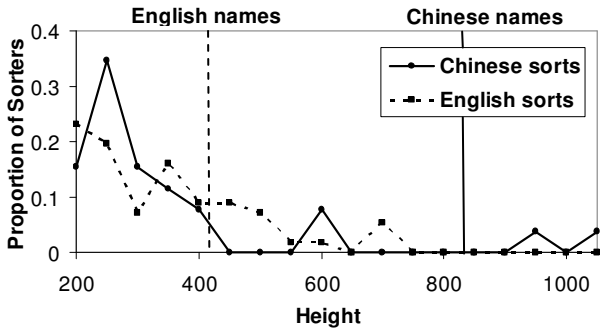

Figure 5: The height (coarse-grainedness) of the Chinese and English naming systems, and sorts produced by speakers of these two languages.

Whereas English was coarser-grained than Dutch in the spatial domain, it is finer-grained than Chinese in the container domain. And the bulk of the sorts produced by speakers of both languages is finer-grained yet. This is consistent with the reasoning proposed above for the apparently privileged status of English in our edit-distance analysis of the container data. Still, as before, we wished to ascertain whether the results are attributable to fine grain *per se*, or to the particular fine-grained partition that English represents. To test this, we also compared the pile-sorts to English-*like* partitions of the container items which are as fine-grained as English but group the items differently – analogously with our creation of Dutch-like partitions of spatial relations, described above. The set of English-like partitions was sampled repeatedly (n=3.5x10^6) by randomly grouping items such that the total number of groups equaled the number of English container terms and the sizes of these groups matched the number of items associated with the English container terms. We then measured the average edit distance from English speakers' sorts to each of these sampled hypothetical English-like partitions (*Min*=41.54, *Mean*=45.67, *Max*=48.02), and the average edit distance from Chinese speakers' sorts to each of these sampled hypothetical English-like partitions (*Min*=44.23, *Mean*=48.01, *Max*=50.31). In both cases the average edit distance of the sorts to actual English (shown in Figure 4) was less than to any of the sampled hypothetical English-like partitions of equally fine grain. This finding suggests that the privileged status of English in our edit-distance results is a consequence not just of its fine-grainedness, but also of the specific groupings of referents that it represents.

Taken as a whole, these reanalyses of the Malt et al. (1999) container data present a picture similar to the one that emerged from our examination of the Khetarpal et al. (2009) spatial data. Similarity judgments as assessed by pile-sorting are fine-grained and quite similar across languages, but also reflect the sorter's native language to some extent. Thus, there is again evidence both for cross-language and for language-specific forces – and thus for both the universalist and Whorfian positions.

Conclusions

Different languages exhibit different systems of semantic categories. It is often assumed that this semantic variation is constrained by, and can be explained by, a universal conceptual space that is partitioned in different ways by different languages. Malt et al. (1999) found evidence consistent with such a language-invariant space, and Khetarpal et al. (2009) assumed such a space existed. In both cases conceptual similarity was assessed through pile-sorting.

We reanalyzed data from these two earlier studies, with a view to reassessing whether pile-sorting on the basis of similarity does or does not reflect language. In both cases we found the same overall picture: pile-sorting was very similar across speakers of different languages (in agreement with the findings and assumptions of the earlier studies), but it also tended to reflect the sorter's native language (in contrast with those studies). Moreover, pile-sorting tended to be semantically finer-grained than any of the languages we considered. These findings suggest several conclusions.

First, they suggest a particular view of the relation of language and thought, namely that: (a) there is a set of fine-grained and potentially cross-cutting conceptual distinctions that may be made, and some languages will happen to mark more of these distinctions than will other languages; (b) distinctions that are unmarked in a language are nonetheless conceptually available to speakers of that language – this is suggested by the fine-grained sorting; and (c) a distinction becomes more salient if it is marked linguistically in one's native language (Hespos & Spelke, 2004) – this is suggested by the effect of language we find. This interpretation is consistent with the general view that "Whorf was half right" and correspondingly half wrong, as has been argued elsewhere (Regier & Kay, 2009).

Second, our results are compatible with the possibility that language may influence cognition in relatively subtle ways that are detectable by some analyses and not by others. Edit distance applied to pile-sorting may be a useful analytical tool, when used in tandem with others, in pursuing this question more generally.

Finally, our results suggest that caution is needed when basing accounts of semantic variation on an ostensibly universal similarity space derived from pile-sorting (e.g. Khetarpal et al., 2009) – because universality cannot be assumed. Similarity judgments are likely to be similar but

not identical across languages, as was the case in our analyses. This highlights an unavoidable tension. A universal conceptual space is a useful theoretical construct for explaining semantic variation, but we have no guarantee that such a thing actually exists – nor, if it does, do we have a completely reliable means of assessing it. Instead, we have somewhat language-colored approximations to such a space, and these should be treated as such. A reasonable treatment may be to average together similarity judgments obtained from speakers of different languages in an attempt to better approximate a universal similarity space, as Khetarpal et al. (2009) did. But any interpretation of results based on such an approximation should be tempered by the awareness that it is merely an approximation.

At the same time, our results leave a number of questions open. The first concerns the contrast between our findings and those of Malt et al. (1999). They found that language was not reflected in sorting by overall similarity, and we found that it was, based on the same data. One possibility, as mentioned above, is that our edit distance analysis is more sensitive than some others, such that it picks up on differences that are missed by other analyses. Is this conclusion correct? Or is our analysis itself inappropriately biased in some respect? Which set of results should be believed? Answering this question is critical to placing our present findings in their proper context.

A second question raised by our findings is the extent to which they generalize to other languages. If we were to examine a new language that partitions semantic space more finely than the languages we have examined here, we would expect to find that pile-sorts produced by people of all backgrounds tend to align more closely with this new fine-grained language than they do with the more coarse-grained languages we have already examined. Is this the case? This question provides a straightforward means of further testing these ideas.

There is also the question of whether these results generalize to other semantic domains. While we have restricted ourselves to the two domains of spatial relations and containers, this was simply a matter of convenience, as the data were readily available. The reasoning behind these ideas however is general in scope, and we would expect to find supporting evidence in other semantic domains as well.

Finally, while these results demonstrate a correlation between language and sorting behavior, they do not demonstrate the causal link claimed by the Whorf hypothesis. It remains an open question whether the observed correlation is attributable to an effect of language on cognition, or to other factors, such as culture influencing both language and cognition.

Regardless of how these questions are eventually answered, we hope that our present initial findings help to make plausible the central idea we have promoted here: a fine-grained conceptual space, largely shared in structure across speakers of different languages, but nonetheless also reflecting the speaker's native language.

Acknowledgments

This work was supported by NSF under grant SBE-0541957, the Spatial Intelligence and Learning Center (SILC).

References

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

Coxon, A. (1999). *Sorting data: Collection and analysis.* Thousand Oaks, CA: Sage Publications.

Croft, W. (2003). *Typology and universals, 2nd edition.* Cambridge, UK: Cambridge University Press.

Deibel, K., Anderson, R., & Anderson, R. (2005). Using edit distance to analyze card sorts. *Expert Systems, 22*, 129-138.

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

Hespos, S. J. & Spelke, E. S. (2004). Conceptual precursors to language. *Nature, 430*, 453 - 456.

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 Conference of the Cognitive Science Society*.

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.

Lucy, J. (1992). *Grammatical categories and cognition: A case study of the linguistic relativity hypothesis.* Cambridge, UK: Cambridge University Press.

Majid, A., Boster, J., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition, 109*, 235-250.

Majid, A., Bowerman, M., Kita, S., Haun, D., & Levinson, S. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences, 8*, 108-114.

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,* 230-262.

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

Regier, T., & Kay, P. (2009). Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences, 13,* 439-446.

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.