Similarity and Variation in the Distribution of Spatial Expressions Across Three Languages

Kristen Johannes (johannes@cogsci.jhu.edu)
Department of Cognitive Science, Krieger Hall
Johns Hopkins University, Baltimore, MD 21218 USA

Jenny Wang (jenny.wang@jhu.edu)
Department of Psychological and Brain Sciences, Ames Hall
Johns Hopkins University, Baltimore, MD 21218 USA

Anna Papafragou (papafragou@psych.udel.edu)
Department of Psychological and Brain Sciences, Wolf Hall
University of Delaware, Newark, DE 19716 USA

Barbara Landau (landau@cogsci.jhu.edu)
Department of Cognitive Science, Krieger Hall
Johns Hopkins University, Baltimore, MD 21218 USA

Abstract

Languages of the world universally encode spatial relationships between objects. However, speakers employ a variety of different language-specific expressions, which may encode culture-specific information about objects and/or different spatial concepts. We ask whether aspects of the encoding of spatial relations across languages nevertheless show common underlying spatial concepts as reflected in the distributions of spatial expressions over spatial sub-types. We examine a set of hypothesized distinctions within the spatial relational concepts of Containment and Support across three typologically distinct languages: English, Hindi, and Mandarin. We find support for two related hypotheses concerning common patterns of variation in (a) speakers' use of select "basic" spatial expressions, and (b) languages' inventory and distribution of expressions across hypothesized Containment and Support subtypes. The results underscore the presence of strong universal similarities in both the extension of basic spatial expressions across relations and in the principles governing the diversity of expressions available for encoding particular relations.

Keywords: Spatial cognition; spatial language; semantics; cross-linguistic analysis

Introduction

All languages have terms for encoding objects as well as a limited set of terms for expressing the spatial relations that hold between objects, such as the English prepositions *in, on,* and *above*. While there is agreement about the universal presence of spatial terms across languages, there is a great deal of debate concerning how these terms encode underlying conceptual-spatial distinctions. Universalist frameworks posit a small number of conceptual-spatial distinctions that are encoded in most languages, with variation primarily centered on which distinctions are encoded in a particular language (e.g. Talmy, 1983; Landau & Jackendoff, 1993). On this view, spatial encoding systems of languages should show strong similarities in the "core" notions that are encoded. A quite different framework, which we will call "culture-dependent" (Levinson et al., 2003; see also Levinson & Wilkins, 2006, Bowerman & Pedersen, 1993), posits that spatial language is built on highly culture-specific notions concerning e.g., culturally salient object properties or configurations, leading to substantial variation across languages that cannot be predicted without a full understanding of the culture. On this view, the spatial encoding systems of typologically diverse language groups should be, at best, weakly related – making contact, for example, in only a handful of relational cases, or following very broad constraints on spatial term extension. In the current work, we examine a set of hypothesized distinctions among spatial relational concepts across three typologically distinct languages, with the goal of assessing whether these languages show any clear similarities in how they encode a broad range of spatial relationships.

Empirical evidence has been adduced for each of the two theoretical frameworks. Looking for variation across languages, Levinson and colleagues (Levinson et al, 2003) and Regier and colleagues (Regier et al., 2013) examined spatial descriptions for a diverse set of spatial scenes included in a battery originally developed by Bowerman and Pederson (1993). Across a large sample of languages (9 languages with 1-36 speakers per sample), these investigators analyzed the spatial term(s) used by the majority of speakers in a language group to encode a given scene. Using multidimensional scaling, Levinson et al. proposed that spatial encoding systems across languages reflect underlying spatial categories that are shaped by a handful of "attractors" – salient spatial scenes that are encoded in similar ways across languages -- and by implicational relationships in the extension of a spatial term to multiple scenes. Similarly, Regier et al. employed an inferential (semantic map) analysis to come to a similar implicational solution, also using the majority, or modal, spatial terms used across languages.

These studies emphasized across-language variation in both the use of basic adpositions and in the complexity of spatial encoding systems. However, they leave open the possibility that deep commonalities exist in spatial language systems across languages, but are reflected in more sensitive measures of language beyond modal expression use. In this paper, we suggest that cross-linguistic commonalities can be captured by measuring within-language variation in expression use – precisely the type of information that is ignored in the analyses above. In particular, we examine two types of within-language variation: the systematic variation in expression use over a range of different spatial cases, and variation in the overall inventory of expressions used by speakers for encoding these different cases.

Strong similarities in spatial concepts across cultures may not be captured by correspondences in the modal spatial term used for specific spatial relation cases within a language. That is, tracking the modal term use across languages may be limited as a measure of cross-linguistic similarity. Here, we examine a different kind of measure: systematicity and variation in the *distribution* of spatial terms across spatial relations. Languages may show surprising consistency as well as systematic variation in the distribution of expressions across many spatial scenes, rather than in the most frequently used expression for a particular scene. This consistency indirectly reflects commonalities in the underlying structure and conceptual organization of spatial relations. Landau and colleagues (Landau et al., under review) pursue this theme in their close examination of the encoding of Containment and Support relations by English and Greek speakers. Despite differences in the overall structure of the spatial language systems in English and Greek, speakers of both languages showed considerable consistency in the way that they extended individual spatial expressions to what were hypothesized to be distinct subtypes of relations. Landau et al. examined the relative frequency with which speakers in each language used language-specific basic locative expressions, (a term borrowed from Levinson, see Levinson et al., 2003; Levinson & Wilkins, 2004) to encode a diverse set of spatial scenes. Speakers' use of the basic expressions *is in/on* in English, and *ine mesa/pano se* in Greek revealed similar patterns of distinctions among subtypes of Containment and Support relations. Furthermore, the spatial descriptions produced by child speakers in each language showed a high degree of similarity to adult-like distributions of the basic expressions for Containment, and systematic differences for Support. Overall, the two languages showed a high degree of consistency in their distribution of a few basic expressions across a range of scenes for the two types of relation. These similarities in the pattern of variability in spatial language use reveals commonalities in the underlying organization of spatial concepts – commonalities that might have been overlooked by analyses that rely on modal expression use. In the current work, we pursue questions of cross-linguistic similarity in not only the distributional characteristics of basic spatial expressions, but also the distributional properties of the spatial inventory, as a whole, across and within languages.

The Current Study

In the current study, we propose a two-part hypothesis concerning systematicity and variation within the broad spatial categories of Containment and Support. We evaluate these hypotheses by comparing the distributions of spatial expressions produced by adult native speakers of three typologically distinct languages – English, Hindi, and Mandarin – in a spatial description task.

Hypothesis 1: Speakers of different languages will show similar distributional patterns of use of their language's basic spatial expressions across different sub-types of Containment and Support relations. To test this, we examine the probabilistic encoding of conceptual structure by language: conceptual differences among spatial relations may be realized as a series of "soft" distributional differences in the use of a small number of expressions[1]. We hypothesize that speakers of different languages will be similar in their probabilistic patterns of use of the basic spatial expression of their language across different spatial relationships. On our proposal, these patterns are predicted to reliably reflect conceptual distinctions represented by speakers in each language. Thus, measuring the distribution of basic expressions in the three languages will reveal commonalities and differences in the conceptual structure underlying spatial expression use in each language.

Hypothesis 2: Speakers of different languages will show similar patterns of variability in the distributional characteristics of all expressions (basic or otherwise) used across different sub-types of spatial relationships. To test this, we examine the variation in the size and distribution of the complete inventory of spatial expressions that are available and used by speakers for encoding different relationships. We hypothesize that variation in the number of expressions used by speakers for different types of relations should also be similar across the three languages.

Methods

[1] We chose our expressions of interest based on agreement among native speakers for encoding the simplest instances of physical containment and support. We consider and test the empirical question of whether the distribution of these basic expressions will reveal common underlying distinctions in the three languages.

Participants and Languages of Interest

Forty-eight adult participants contributed to the corpus of spatial descriptions: 16 each of native English speakers, native Mandarin speakers, and native Hindi speakers. English and Mandarin descriptions were collected in person and online from speakers in the Johns Hopkins University community. Hindi descriptions were collected in person from the Baltimore, MD and Boston, MA areas. A profile of each of the three languages and their respective spatial systems is presented in Table 1, below. The corpus of English descriptions was originally collected by Landau et al. (under review) as part of a larger cross-linguistic data set comparing adult and child use of basic spatial expressions in English and Greek speakers. In the current study, we use the same proportions of and analysis of English basic spatial expressions initially reported in Landau et al. To this we add a new corpus of Hindi and Mandarin spatial descriptions and analysis of basic expression use in these languages. Finally, we also examine properties of the complete spatial expression inventory for each language, an analysis that is new to all three language corpora.

Materials and Procedure

Spatial Relations Battery The study employed a Spatial Battery of 80 static scenes, developed by Landau and Papafragou (Landau et al., under review). Each scene depicted a spatial relation between two objects: a figure object indicated by an arrow, and a ground object outlined in white. The spatial relations of interest in the current study were derived from two broad categories: Containment and Support, which encompassed 44 of the 80 scenes, further articulated into a structure of subcategories. Containment was divided into 6 subtypes and Support was divided into 5 subtypes. Each subtype featured 4 scenes. Figures 1A and 1B present the respective Containment and Support subtypes, with a single exemplar from each.

Spatial Description Task and Coding Participants were asked to provide a description in their native language for each scene, as an answer to the question "Where is the [object indicated by the arrow]?" Each participant was presented with the 44 scenes in a different random order, and descriptions were transcribed and coded for spatial content, which differed in form depending on the language. As part of the transcription and coding process, we defined separate basic spatial expressions (following the work of Levinson et al., 2003[2]), listed in Table 1, used to describe Containment and Support relations for each language.

[2] Specifically, Levinson et al. (2003; see also Levinson & Wilkins, 2006) define the "basic locative construction/expression" (BLC) in a language as the most natural response to the question of "where is the (figure object) located?", acknowledging that other expressions might be used in certain pragmatic contexts.

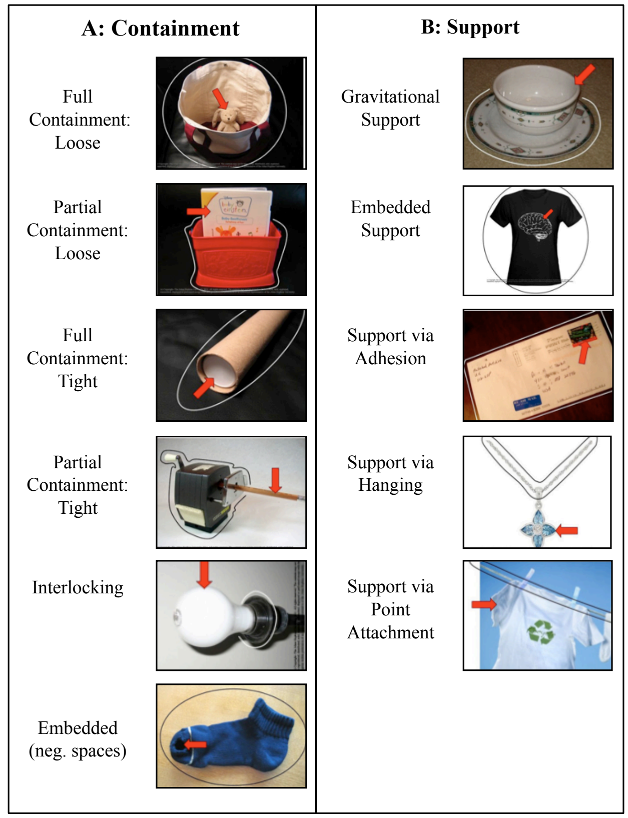

Figure 1: Containment (A) and Support (B) subtypes and examples. Each sub-type included 4 examples.

Table 1: Language profiles and spatial expressions for languages of interest.

	English	Hindi	Mandarin
Family	Germanic	Indo-Iranian	Sino-Tibetan
Word Order	SVO	SOV	SVO
Spatial Expression Form	Verb + Preposition	Postposition + Verb	(Verb+) Prepostion + Loc.Noun
Basic Expression: Containment	is_V $in(side)_P$ be in	$main_P$ hai_V in be	zai_P $li\text{-}_{LocN}$ locate-at interior
Basic Expression: Support	is_V $on(top)_P$ be on	par_P $(rakha_V)$ hai_V on (place) be	zai_P $shang_{LocN}$ locate-at top

Analyses and Results

Hypothesis 1: Systematicity in Basic Expression Use

Analysis We examined the distribution of basic spatial expressions in each language (see Table 1 for the basic expressions measured). To measure this, we calculated the proportion of descriptions for each subtype of Containment and Support for which speakers used the basic expression in their language. We then used logistic regression analyses to determine whether differences among subtypes of relations predicted reliable differences in rates of use of the basic expression by speakers of a given language. These models were conducted separately within each language for Containment and Support scenes. Subtypes were tested against one another by coding the weights in the regression algorithm to reflect a set of orthogonal contrasts between subtypes of relations in the battery, allowing us to interpret regression coefficient values as differences in language use for dimensions that distinguished subcategories of relations from one another.

Results Speakers of all three languages showed variability in the use of their language-specific basic spatial expression across sub-types of Containment (Figure 2) and Support (Figure 3). We looked for similarity in the patterns of this variability – the way in which speakers distributed their use of the basic spatial expression – across the three language groups. The similarity of these patterns was statistically confirmed through mixed model logistic regression analyses, conducted separately for descriptions of Containment (Table 2) and Support (Table 3) relations in each of the three languages.

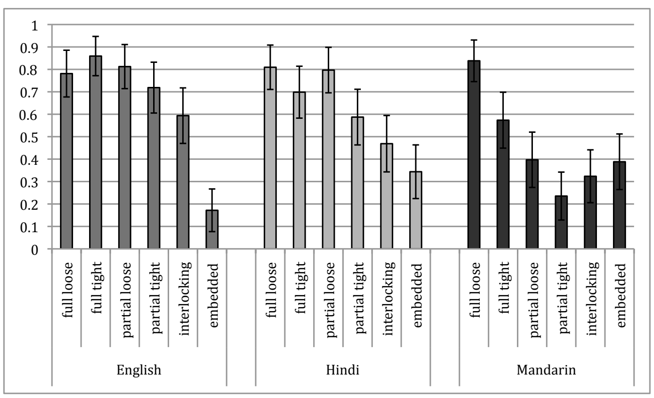

Figure 2: Proportion of basic expression use grouped by subtype for Containment in English (*is in*), Hindi (*main hai*), and Mandarin (*zai li-*).

For Containment relations (Figure 2, Table 2), speakers of all three languages showed common distributional differences for 2/4 possible hypothesized conceptual distinctions: speakers were more likely to use their respective basic spatial expressions – *is in, main hai,* and *zai li-* – to describe Full- and Partial- Containment relations compared to Interlocking and Embedded relations, and used the basic expression at different rates for Interlocking compared to Embedded relations. Additionally, speakers of two languages, Hindi and Mandarin, distinguished Loose-fitting from Tight-fitting Containment relations, and Mandarin speakers were more likely to use their basic spatial expression *zai li-* for Full- compared to Partial-Containment sub-types.

Table 2: Model coefficient estimates (***β***) and statistical reliability (p = Pr(z)) for logistic regression analysis of basic expression use for Containment in each language.

	English		Hindi		Mandarin	
	β	p	β	p	β	p
Intercept	**0.86**	**<.01**	0.79	*ns*	0.10	*ns*
Full/Part Contain v Interlock/Embed	**3.18**	**<.01**	**2.86**	**<.01**	**2.07**	**<.01**
Contain: Full v. Partial	0.41	*ns*	0.43	*ns*	**2.18**	**<.01**
Contain: Loose v. Tight	-0.01	*ns*	**-1.29**	**<.05**	**-1.31**	**<.05**
Interlocking v. Embedded	**-1.17**	**<.01**	**-0.75**	**<.05**	**0.68**	**<.05**

Likewise, for Support relations (Figure 3) speakers in each language tended to use their basic spatial expression at greatest rates for Gravitational Support relations and at increasingly lower rates for Embedded Support, followed by Support via Adhesion, Hanging, and Point Attachment.

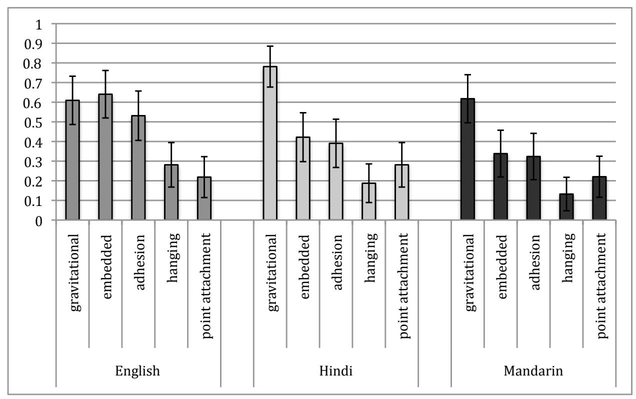

Figure 3: Proportion of basic expression use grouped by subtype for Support in English (*is on*), Hindi (*par (rakha) hai*), and Mandarin (*zai shang-*).

However, regression analyses (Table 3, below) make clear that speakers of each language were reliably more likely to use their basic spatial expressions – *is on, par (rahka) hai* and *zai shang-* for English, Hindi, and Mandarin, respectively – to express Gravitational Support relations compared to sub-types of Support via Hanging, and Point Attachment. And that speakers of Hindi and Mandarin, but not English, further distinguished between Gravitational Support relations compared to sub-types of Embedded Support and Support via Adhesion. Our analyses show that the variability in basic expression use *within* languages reveals both similarities and differences in the patterns of probabilistic encoding of Containment and Support relations *across* languages. All three languages share a set of reliable distinctions between sub-types, but speakers of some languages make a greater number of distinctions (reflecting, perhaps, more fine-grained encoding) than speakers of others.

Table 3: Model coefficient estimates (***β***) and statistical reliability for logistic regression analysis of basic expression use for Support in each language.

	English		Hindi		Mandarin	
	β	***p***	***β***	***p***	***β***	***p***
Intercept	0.72	*ns*	-0.66	*ns*	**-1.24**	**<*.05***
Gravitational v. Embedded	0.21	*ns*	**-1.43**	**<.01**	**-1.90**	**<.01**
Gravitational v. Adhesion	-0.45	*ns*	**-1.57**	**<.01**	**-1.97**	**<.01**
Gravitational v. Hanging	**-2.10**	**<.01**	**-1.47**	**<.01**	**-1.60**	**<.01**
Gravitational v. Pt attach	**-2.58**	**<.01**	**-0.81**	**<*.05***	**-0.89**	**<*.05***

Hypothesis 2: Variation in the Complete Expression Inventory

Analysis The task of finding translational equivalents and evaluating variation across all three languages becomes increasingly difficult beyond the basic spatial expressions. Thus, we sought to measure the general variation in the inventories of expressions used to express Containment and Support relations in each language. To do this, we calculated the entropy, H(X), associated with the set X of expressions produced for each scene:

$$H(X) = -\sum_i P(x_i)\log P(x_i)$$

Higher values for H reflect increasingly uniform distributions of expression use (many equally-probable expressions), while low values for H reflect distributions with few high-frequency descriptions. We used these values as indirect measures of the variability of spatial descriptions in each language and conducted correlations to establish whether the variability in the distribution of descriptions across scenes was similar across the three languages.

Results The complete inventories of (both basic and other) spatial expressions for each language also varied over subtypes, with some subtypes eliciting only a few different expressions and others eliciting a wide variety. The mosaic of plots in Figures 4 (Containment sub-types) and Figure 5 (Support sub-types) give a snapshot of frequency and distribution of different spatial expressions across different relation sub-types in each language, and each pie piece represents the frequency with which a single expression[3] was used in the language. To determine whether different languages were similar in the degree to which speakers' expression choice varied across different spatial relations, we calculated entropy values for each Containment and Support scene in each language and computed correlations (separately for Containment and Support scenes) for each pair of languages.

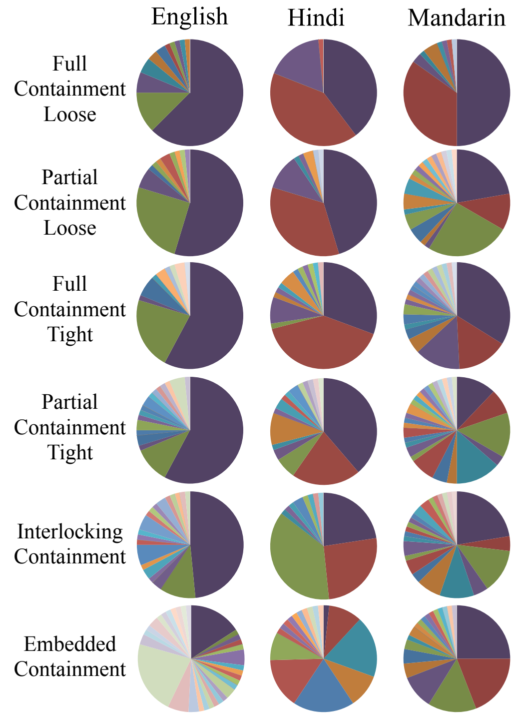

Figure 4: Distribution of spatial expressions for Containment sub-types in each language. Expressions within a language are consistently colored across sub-types and the basic expression is the same color (dark purple) within and across languages.

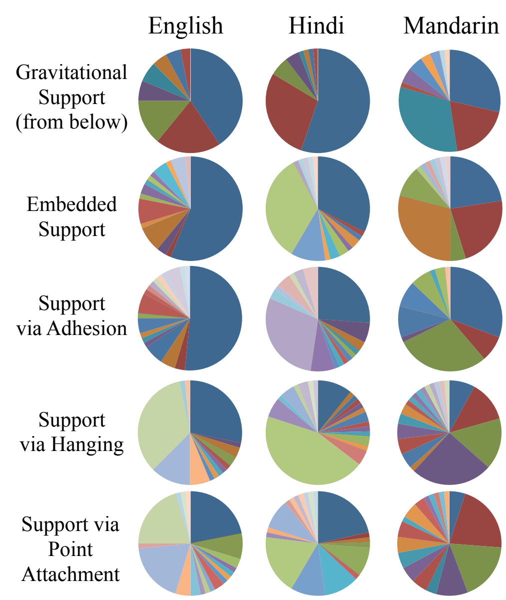

Figure 5: Distribution of spatial expressions for Support sub-types in each language. Expressions within a language are consistently colored across sub-types and the basic expression is the same color (medium blue) within and across languages.

[3] We differentiated expressions on the basis of adposition, verb, spatial noun and/or locative marker. For example, seemingly related prespositions such as English *in, inside,* and *within,* were treated as distinct expressions.

Entropy values were computed across scene (44 scenes total: 24 Containment, 20 Support) for the spatial descriptions in each language. These values reflect both information about the frequency and variability with which expressions were used for each scene. We entered these values into a series of Spearman rank-order correlation analyses (Table 4, Spearman's ρ significance established via randomization tests with 10,000 permutations each), to measure the strength of correlation between entropy (H) values for pairs of languages.

Table 4: Spearman rank-order correlations (ρ) between entropy (H) values across Containment and Support relations (by item) for pairs of languages.

	English-Hindi	*English-Mandarin*	*Hindi-Mandarin*
Containment Correlations	**ρ = 0.62**, *p<.01*	**ρ = 0.56**, *p<.05*	**ρ = 0.76**, *p<.01*
Support Correlations	**ρ = 0.54**, *p<.05*	**ρ = 0.50**, *p<.05*	**ρ = 0.46**, *p<.05*
All Items Correlations	**ρ = 0.44**, *p<.01*	**ρ = 0.53**, *p<.01*	**ρ = 0.32**, *p<.05*

Languages showed moderate-to-high positive relationships (ρ = 0.32-0.76) in their tendency to distribute spatial descriptions for different spatial scenes across few (low entropy) vs. many (high entropy) expressions. Correlations among individual scenes were reliable across all language pairs, both when computed separately for Containment and Support relations and when collapsed over category[4].

Conclusions

In the current study we used a novel set of distributional analyses to uncover deep commonalities, as well as differences[5], in the structure of spatial encoding systems across three typologically distinct languages. Our first analysis uncovered consistency in the distribution of basic spatial expressions across a wide range of relations in each language: speakers of each language show similar tendencies to use a single basic expression at different rates to distinguish among subtypes of Containment and Support relations. Our second analysis extended this cross-linguistic consistency to variation in the spatial expression inventory across languages: languages were correlated in the extent to which speakers preferred to use either few expressions, each used frequently, or many low-frequency expressions to encode a particular spatial relation scene. These results come together to inform our view of cross-linguistic encoding of spatial relationships in which similarities arise as a function of systematic variability in speakers' use of (a) basic spatial expressions over sub-types of relations; and (b) the complete inventory of expressions available within a language for encoding different relations.

The current work highlights universality in both the extension of basic spatial expressions across relations and in the principles governing the diversity of expressions available for encoding particular relations. Similarities across languages arise despite clear typological differences in the origins and structure of the sampled languages as well as possible widespread culture-specific pressures on the expression of information about objects and location. We hope, in future work, to extend this analysis to language samples with even greater typological diversity.

Acknowledgments

Thanks to Sonal Chatter and Mengxi Yuan for assistance and to Colin Wilson and three anonymous reviewers for helpful comments. A.P. acknowledges partial support from NSF Grant BCS-0749870.

References

Bowerman, M., & Pedersen, E. (1992). Topological relations picture series. In Stephen C. Levinson (Ed.) *Space stimuli kit 1.2,* 51. Nijmegen: Max Planck Inst. for Psycholinguistics.

Bowerman, M., & Pedersen, E. (1993). Cross-linguistic studies of spatial semantic organization. In *Annual Report of the Max Planck Institute for Psycholinguistics 1992,* 53-56.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences, 16,* 217-238.

Landau, B., Johannes, K., Skordos, D., & Papafragou, A. (under review). Universality and variation in the representation of Containment and Support: Evidence from English- and Greek-speaking children and adults.

Levinson, S.C., Meira, S, et al. (2003). 'Natural concepts' in the spatial topological domain – adpositional meanings in cross-linguistic perspective: An exercise in semantic typolgogy. *Language, 79,* 485-516.

Levinson, S.C., & Wilkins, D.P. (Eds.) (2006). *Grammars of space: Explorations in cognitive diversity.* Cambridge: Cambridge University Press.

Regier, T., Kay, P. & Khetarpahl, N. (2007). Color naming reflects optimal partitions of color space. *PNAS, 101,* 1436-1444.

Regier, T., Khetarpahl, N., & Majid, A. (2013). Inferring semantic maps. *Linguistic Typology, 17,* 89-105.

Talmy, L. (1983). How language structures space. In H.L. Pick Jr. and L.P. Acredolo (Eds.) *Spatial Orientation: Theory, Research, and Application*. New York: Plenum.

[4] There may be other appropriate standards for comparing similarity across languages. In their work on color naming, for example Regier et al. (2007) propose a similarity metric and evaluate its utility by "shifting" the space in arbitrary yet semantically contiguous ways. Spatial categories, however, lack the semantic contiguity present in the color space, making the Regier et al. approach and, specifically, shifted standard of comparison, less viable for analyses of spatial encoding.

[5] An anonymous reviewer raised a concern that our hypotheses are (un)falsifiable. However, on our account falsifiability is possible, specifically predicting the failure to find the same sets of relative differences in basic expression use (H 1) and failure to find correlations between distributions of expressions (H 2).