Inferring Conceptual Structure from Cross-Language Data

Terry Regier[*] (terry.regier@berkeley.edu)
Naveen Khetarpal[†] (khetarpal@uchicago.edu)
Asifa Majid[|] (asifa.majid@mpi.nl)

[*] Department of Linguistics, Cognitive Science Program, UC Berkeley
[†] Department of Psychology, University of Chicago
[|] Max Planck Institute for Psycholinguistics, Nijmegen

Keywords: semantic universals; conceptual structure; language and thought; semantic maps.

Semantic categories vary across languages, but this variation is constrained: many logically possible semantic categories are not attested, and similar categories often appear in unrelated languages. This pattern suggests a universal conceptual basis to the variation, such that different languages provide different snapshots of the same conceptual terrain. A *semantic map* is a means of capturing this idea. Formally, a semantic map is a graph in which vertices represent different semantic uses or functions, and edges connect closely related uses. It is assumed that the graph structure is universal, and that language-specific categories always pick out *connected subgraphs* of the universal graph. A semantic map thus compactly represents what patterns of variation one may and may not expect to find in a given semantic domain, in terms of presumptively universal conceptual structure. Semantic maps have been widely used; for recent reviews see Haspelmath (2003), Croft (2003), and Cysouw et al. (2010) plus other papers in the same volume.

The structure of the graph is generally determined by a cross-language dataset of semantic categories: edges are added so that each category in the dataset is connected. The task of constructing a semantic map in this manner is generally done by hand, and the task can be time-consuming and potentially error-prone. It would therefore be useful to automate this process. However the computational problem of inducing the optimal semantic map from cross-language data has not previously been formally addressed. Croft and Poole (2008: 6-7) conjectured that this problem may be computationally intractable, and they considered this potential intractability to be a shortcoming of semantic maps as a representational tool in semantic typology. Here, we address this issue by casting the *semantic map induction* problem in formal terms. We note that this problem is formally identical to a problem that superficially appears unrelated: inferring a social network from disease outbreaks in a population. Angluin et al. (2010) have recently shown that this social network inference problem is computationally intractable (specifically: hard to approximate, e.g. Trevisan, 2004; Vazirani, 2001, ch. 29), but that an efficient algorithm exists that approximates the optimal solution nearly as well as is theoretically possible. It follows that both the computational intractability and the applicability of the approximation algorithm hold of semantic map induction. We apply this algorithm to the cross-language data of Haspelmath (1997) on indefinite pronouns, and obtain a simpler map than the published one. We apply the same algorithm to the data of Levinson et al. (2003) on spatial categories, and obtain a map that accommodates categories from a new language, as novel connected subgraphs of the induced graph structure. We conclude that presumptively universal conceptual structure consistent with cross-language data can be straightforwardly inferred, that the issue of computational intractability, while real, need not deter researchers, and that formalization of problems in semantic typology can highlight useful connections to structurally related problems elsewhere.

Acknowledgments

We thank Martin Haspelmath, Stephen Levinson, and Sérgio Meira for sharing their data. This work was supported by NSF under grant SBE-0541957, the Spatial Intelligence and Learning Center (SILC).

References

Angluin, D. et al. (2010). Inferring social networks from outbreaks. In Hutter, M. et al. (Eds.), *Algorithmic Learning Theory 2010, LNCS 6331* (pp. 104-118). Berlin: Springer.

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

Croft, W. & Poole, K.T. (2008). Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics 34*, 1–37.

Cysouw, M. et al. (2010). Introduction to the special issue "Semantic maps: Methods and applications". *Linguistic Discovery, 8*.

Haspelmath, M. (1997). *Indefinite pronouns*. Oxford.

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

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

Trevisan, L. (2004). Inapproximability of combinatorial optimization problems. Technical report TR04-065, Electronic Colloquium on Computational Complexity.

Vazirani, V. (2001). *Approximation algorithms*. Springer.