

Quantifying semantic regularity across languages

Asifa Majid & Stephen C. Levinson

Semantic maps in typological work are often produced on the basis of underlying conceptual spaces constructed by intuition and inspection. Here we argue that if the underlying conceptual spaces are thought of as multi-dimensional spaces, structured as similarity spaces, then this allows the application of sophisticated quantitative methods. For the typologist interested in quantifying how similar semantic categorization is across languages, these methods offer exciting new possibilities. Focussing on specific domains, they can be used to study different - although interrelated - questions, such as:

(1) What are the semantic distinctions being made within a domain? And where languages make different distinctions, do they respect the coherence of the same underlying conceptual space?

(2) How similar (or different) are languages in their overall pattern of categorization? Can we measure degrees of convergence?

(3) Are children faster at learning semantic distinctions that are typologically frequent in comparison to rare distinctions?

Drawing on findings from several collaborative cross-linguistic projects based at the Max Planck Institute for Psycholinguistics, we examine the semantic categorization of events as reflected in verbs and constructions, and highlight techniques for analyzing large cross-linguistic data sets that can capture both shared category structure and language variability.

In the case of event domains, the projects begin with an etic grid of event types - a set of videoclips - which vary along a number of parameters pertinent to the domain of study. Speaker descriptions are then elicited from a range of geographically, genetically and typologically diverse languages. The descriptions are analyzed using multivariate statistics, such as factor analysis, correspondence analysis and multidimensional scaling. These techniques not only visually represent cross-linguistic regularity, but also quantify precisely how much structure is shared, and what constitutes an unusual pattern of categorization.

The picture emerging is one of robust regularity. For instance, for "cutting and breaking" events all languages recognize a dimension having to do with how predictable the location of separation in an entity will be. Categorization of reciprocal events shows much more variation, with some quite different solutions to the problem of how to encode such events, but nevertheless recurrent categorization patterns emerge. Finally, we show how the multivariate conceptual spaces extracted from cross-linguistic studies can be used to investigate language acquisition too.