Generalizing Language Comparison

Michael Cysouw

Max Planck Institute for Evolutionary Anthropology, Leipzig

The method proposed in the target article by Croft & Poole (henceforth C&P) is an important step towards the generalization of semantic map-like approaches as currently used in linguistic typology. However, I think we can even go further with this methodological generalization. In this comment, I would like to quickly sketch what I think could be seen as an even more overarching approach to typology (or language comparison in general), of which C&P's method is just one possibility. This does not mean that C&P's method is wrong in any way, but I would like to suggest that their Optimal Classification Nonparametric Unfolding algorithm is a method that is fine for some kind of data, but will not be the best choice in all situations.

My main objective to the method proposed by C&P is that their algorithm directly jumps from a particular kind of data (binary coded language-particular constructions) to a visual representation (a two-dimensional display), which is then left to the researcher to interpret. Although such an all-in-one method has the prospect to allow also non-mathematically oriented linguists to easily make use of such methods (when encapsulated in a proper user interface), the downside is that it is not possible to specify more details. And, as far as I can see, this is a principle limit of C&P's method: it does no appear to be easily extendable. A more general approach, that I will advocate, implies that there are many more decision needed on the side of the researcher, which presupposes knowledge of the (often rather technical) issues involved. However, many of these issues could easily be given a default setting so the algorithm would revert to something like the C&P method if no individual specification are made.

In the most general sense, I think language typology consists of establishing a *pairwise similarity* between a collection of entities investigated. These entities could either be a set of languages, and the similarity between the languages is established by comparing a selection of comparative concepts, or these entities can be a set of comparative concepts, and the similarity between these characteristics is established by comparing their expression in a selection of languages. Both approaches are methodologically largely equivalent—by exchanging 'languages' for 'comparative concepts', and *vice versa*—barring of course various differences in the selection of relevant data and in the interpretation of the results. In this short note, I will restrict myself to the question of the relation between comparative concepts, which is also the problem addressed by C&P.[1]

Given a set of comparative concepts of interest to a linguist, cross-linguistic research approaches these concepts by investigating their expression in a wide array of languages. The central assumption is that recurrent similar expression of two concepts among geographically and genealogically unrelated languages indicates that these concepts have some kind of semantic, functional, cognitive or structural similarity. The choice among these (and other) possible interpretations is a difficult issue which I will not further deal with here. Until here I think there is no disagreement between me and C&P. Where I find their approach lacking is with:

- the restriction to binary "yes/no" coding of applicability of the language-particular

[1] The term 'comparative concept' used here is supposed to remain agnostic about whether these are conceived of as functions, meanings, extensions, contexts, features, or even abstract statistical properties of stretches of text. The only important assumption is that comparative concepts should be formulated generally enough to be applicable to every human language. I thank M. Haspelmath (p.c.) for

constructions to a comparative concept;

- the missing possibility to specify the relation between the different constructions from within the same language;
- the restriction to one kind of feedback to the researcher about the internal structure of the data (i.e. the graphical display).

First, it is unfortunate that it is not possible to specify (if one so wishes) that a particular construction is either commonly, or only sparingly, used to express a particular comparative concept. So, instead of the binary yes/no coding used by C&P it should be possible to express this as a continuum. Note that very often there are structural reasons that explain why there are differences between the frequency of usage of a particular construction. However, such an insight can only be expressed typologically by adding extra comparative concepts that distinguish the structural factors.

Second, in the kind of research design as discussed by C&P there is normally more than one construction per language relevant. In the algorithm of C&P, all

suggesting this term.

constructions are weighted equally, also if there would be, for example, two constructions from language A and ten constructions from language B in the dataset. This would, without correction, give a much greater weight to the structure of language B compared to language A in the comparison. It would be good to at least to have the possibility to check whether such an implicit decision is of any relevance to the results (and I expect that such weighting of the input would be possible to implement in the C&P method—if it is not already available). More problematic is the assumption of C&P that it is possible to neatly distinguish the different constructions from within a language. My impression is that it is often not easy to decide whether a set of comparative concepts is expressed by some closely similar constructions, or whether these constructions should all be treated as the same one.

It is possible to address these issues by adding an extra step in the analysis of the cross-linguistic data. I would like to call this level of analysis the LANGUAGE-SPECIFIC PERSPECTIVE on the comparative concepts. The idea is that at this stage a summary is given about the way a particular language addresses the relation between the comparative concepts under investigation, combining all language-particular constructions and the relations between them. In its most general expression, this language-specific perspective takes the form of a similarity matrix of pairwise similarities between all pairs of concepts, just on the basis of the data from one single language. To establish such a matrix, various decision have to be taken how to evaluate the coding similarity between two concepts in each language. However, it is important to realize that all of these decision are language particular, meaning that the argumentation does not have to be the same for all language in the sample.

The next step of the comparison is to combine the language-specific perspective into a cross-linguistic perspective. Because the language-specific perspectives are already expressed in a standardized form (a similarity matrix of concepts by concepts), this combination amounts simply to adding up all language-specific matrices together (with the possibility to weight the individual matrices to counter expected biases in the language sample). The resulting summed similarity matrix represents the cross-linguistic view on the relation between the comparative concepts under investigation.

The final step of this general approach to language comparison is then the interpretation of such a matrix of pairwise similarities. C&P use a geographical map-like graphical display to help a human being to make sense of a large set of numbers. This is a well-known approach from (related) methods like multidimensional scaling, principal components analysis or correspondence analysis. However, an enormous amount of methods to make sense of similarity matrices has been developed in recent decades, using for example tree structures (e.g. hierarchical clustering), tree-like networks (e.g. splits graphs), graph display algorithms (e.g. force-based graphs), or cluster analyses (e.g. k-means). They all show slightly different perspectives on the data, so it is often very helpful to try out more than one approach before any conclusions are drawn. More so, in many of these approaches there are measures developed, accompanying the analysis, that indicate how significant a particular display is. This is important, because one central problem with all of the above displays is (including the one from C&P's method) that they will always show something, and that something might even look interesting to a human eye, also if the structure shown is just a semi-random artifact of the method applied.

This short survey of an even more general approach to language comparison glosses over many of the (important) details, but I hope to have made clear that I think the proposals by C&P is a large step in the right direction—a direction that should be taken even further.