Cross-linguistic comparison of complexity measures in phonological systems

Steven Moran
University of Zurich
Damian Blasi
Max Planck Institute

Although complexity of subsystems varies greatly across languages, the compensation hypothesis states that if a language's structure is complex in one area, it will simplify in another (e.g. Martinet 1955, Hockett 1955, Aitchison 2000). An assumed truism is that these differences balance out cross-linguistically, so that all languages tend to be equally complex (e.g. Hockett 1958, Akmajian et al. 1979, Crystal 1987, McMahon 1994, Dixon 1997). This belief is furthered by the long-held view that linguistic structures are not affected by geographic or societal factors, with vocabulary being an exception, e.g. Sapir 1912.

Recently, assumptions about equal language complexity have been challenged (e.g. McWhorter 2001, Kusters 2003, Dahl 2004, Hawkins 2004, Shosted 2006, Miestamo et al. 2008, Givón & Shibatani 2009, Sampson et al. 2009, Sinnemaeki 2011), including findings of correlations between complexity and geographic or sociocultural settings (e.g. Perkins 1992, McWhorter 2001, Kusters 2003, Trudgill 2004, Hay & Bauer 2007, Nichols 2009, Lupyan & Dale 2010).

In this paper we examine whether or not different aspects of phonological systems correlate with each other and whether or not there are trends that correlate with non-linguistic factors, such as geography and population. For this work, we have compiled together several phonological typology databases. The compiled dataset includes segment inventory data (Crothers et al, Maddieson 1984, Maddieson & Precoda 1990, Moran 2012), distinctive features (Hayes 2009, Moran 2012), syllable structure (Maddieson 2011), rhythm types (Goedemans & van der Hulst 2011) and phonological word domains (Bickel et al. 2009). The data are augmented with information about languages' genealogy (language family stock and genus; Lewis 2009, Dryer 2011), geography (regions and geo-coordinates; Haspelmath et al. 2011) and demography (population; Lewis 2009).

Our analyses reveal that, when all languages are taken into account, many (little though significative) patterns concerning phonological and non-phonological data emerge. Examples include correlations between the size of consonant and vowel inventories (Spearman's ρ: 0.12, $p < .00001$), size of the segment inventory and mean word length (Spearman's ρ: -0.309, $p < .00001$), population size (Spearman's ρ: 0.331, $p < .00001$) or latitude (Spearman's ρ: 0.195, $p < .00001$). Conspicuously, when linguistic family is included as a variable, we find no homogeneous picture: some families follow the general correlations, whereas others do not or even go against it. This strong dependency of the patterns with genealogy is even present in the statistical distributions of phonological repertoires, which leads to general cross-linguistic distributions with non-trivial (i.e., parametric) structure. We explore the consequences these results have for our

understanding of the explanatory role correlations have, and the time scales in which phonological change occurs.

As a corollary, we discuss the connections between rarity (i.e., cross-linguistic frequency) and complexity and the evidence from cognitive sciences in this direction. Finally, we move into unexplored territory and present results of cross-linguistic statistical aspects of distinctive features.