Treebanks and Evolutionary Simulation for Explaining Typological Patterns

Robert Malouf
San Diego State University
Department of Linguistics & Asian / Middle Eastern Languages
E-mail: rmalouf@mail.sdsu.edu

Abstract

Recent work in Evolutionary Phonology (Blevins 2005, 2006, Blevins & Wedel 2008, Yu 2007, among others) has developed alternate explanations for typological universals or tendencies found across the sound systems of unrelated languages. This research emphasizes the role of patterns of language use and language change in the development of cross-linguistic patterns, rather than placing the burden of explanation on synchronic cognitive factors (i.e., Universal Grammar).

In this talk, we will review extensions of this work into the domain of morphology (Ackerman, Blevins, and Malouf to appear). We investigate the Paradigm Cell Filling Problem, a particular question for inflectional morphological systems which has received relatively little attention in the theoretical literature. Specifically, we ask: How do speakers of morphologically complex languages predict the full inflectional (or derivational) paradigms of novel words, given exposure to a small number of surface word forms? For example, a noun in Tundra Nenets can appear in 210 different inflected forms. Given exposure to one of these forms of a novel noun, how does a Tundra Nenets speaker predict the other 209 forms?

Our hypothesis is that speakers' need to solve the Paradigm Cell Filling Problem serves as a strong evolutionary pressure on language, which in turn leads morphological systems to develop in particular directions. Thus the Paradigm Cell Filling Problem is an indirect explanation for some of the typological patterns found cross-linguistically in morphological systems. In order to test our hypotheses about language development, we perform computer simulations of language evolution across many generations, to see which factors cause which patterns to arise. In this way, we treat language as a complex adaptive system and therefore link linguistic study to larger trends in the biological and social sciences (e.g., Miller and Page 2007).

As with many other complex adaptive systems, the outcome of our simulations of linguistic evolution can be highly dependent on the initial condi-

tions. Therefore, to be reliable, our simulations need to be based on detailed and accurate information about synchronic language states. As the domain of investigation moves from morphophonology to morphosyntax, it becomes more difficult to find the information we need in conventional typological databases and lexicons. The kind of information we need can only be found in treebanks – corpora with highly detailed linguistic annotations. Therefore, the development of large, richly annotated corpora in a variety of typological diverse languages is crucial to the evolutionary program for explaining cross-linguistic evolutionary patterns.