A GRADUAL PATH TO HIERARCHICAL PHRASE-STRUCTURE: INSIGHTS FROM MODELING AND CORPUS-DATA

WILLEM ZUIDEMA

Institute for Logic, Language and Computation, University of Amsterdam,
Plantage Muidergracht 24, 1018 HG, Amsterdam, the Netherlands
jzuidema@science.uva.nl'

Neo-darwinian scenario's for the evolutionary emergence of a trait involve some sort of characterisation of (i) the set of possible phenotypes-genotypes that can be reasonably assumed to have been "available" (the *strategy set*); (ii) the consequences any of these phenotypes-genotypes for survival and reproduction (the *fitness function*), and (iii) a path of ever-increasing fitness that leads from phenotypes without the trait to phenotypes with it (Parker & Maynard Smith, 1990). This suggests an agenda for scientific research into the evolution of language, which consists of first identifying the relevant traits that are in need of an evolutionary scenario, and then filling in the details in each of these three domains, for each of these traits.

With (Jackendoff, 2002) and many others, we have argued that three traits in need of a detailed evolutionary scenario, are combinatorial phonology, compositional semantics and hierarchical phrase-structure (Zuidema, 2005). Only with such detailed scenarios in place, can we start evaluating the relative plausibility of various assumptions on heritability, selection pressures and the role of epigenesis and cultural evolution. In this paper we will focus on the evolution of hierarchical phrase-structure, for which no such detailed neo-Darwinian model has been proposed yet – surprisingly, perhaps, given its centrality in many debates. We will briefly discuss why some existing models are incomplete from this perspective, and argue that recent work on exemplar-based grammar provides a solution to the problems encountered when trying to extend them.

Our argument consists of two steps. First, we consider syntax in modern language. We demonstrate that exemplar-based models can do justice to the unlimited productivity of natural language (e.g. (Bod, 2006)). However, we also observe that many regularities in natural languages that look systematic, are in fact stored holistically in the memory of language users (pseudo-productivity). Using corpus data and the statistical techniques developed in (Zuidema, 2007), we show that the storage in memory of larger fragments of language is the rule rather than the exception. We conclude that cognitively adequate formalisms for syntax

must allow for productive units of almost any size: from single morphemes and context-free rules of combination, to complete holophrases.

Second, we consider the evolutionary history of syntax. We show that one such formalism – *probabilistic tree substitution grammars*, used with success in computational linguistics – allows for a very natural way to define a fitness function, which in turn allows us to show an evolutionary path from a communication system without hierarchical structure to one which, like natural language, shows hierarchical phrase-structure and full productivity along side pseudo-productivity and a heterogeneous store of productive units.

The resulting model is the first precise, neo-Darwinist model that shows a gradual route to hierarchical phrase-structure; however we will highlight relations with less formal proposals from research on *construction grammar* (e.g. (Verhagen, 2005)) and *formulaic language* (e.g. (Wray, 2000)), and formal, but not neo-Darwinist models such as Data-Oriented Parsing (Bod, Scha, & Sima'an, 2003) and Batali's negotiation model (Batali, 2002).

References

Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: formal and computational models*. Cambridge, UK: Cambridge University Press.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*(23). ((Special Issue on Exemplar-Based Models in Linguistics))

Bod, R., Scha, R., & Sima'an, K. (Eds.). (2003). *Data-oriented parsing*. Chicago, IL: CSLI Publications, University of Chicago Press.

Jackendoff, R. (2002). *Foundations of language*. Oxford, UK: Oxford University Press.

Parker, G. A., & Maynard Smith, J. (1990). Optimality theory in evolutionary biology. *Nature*, *348*, 27-33.

Verhagen, A. (2005). *Constructions of intersubjectivity. discourse, syntax, and cognition.* Oxford: Oxford University Press.

Wray, A. (2000). Holistic utterances in protolanguage: The link from primates to humans. In J. R. H. Chris Knight & M. Studdert-Kennedy (Eds.), *The evolutionary emergence of language: Social function and the origins of linguistic form.* Cambridge: Cambridge University Press.

Zuidema, W. (2005). *The major transitions in the evolution of language*. Unpublished doctoral dissertation, Theoretical and Applied Linguistics, University of Edinburgh.

Zuidema, W. (2007). Parsimonious Data-Oriented Parsing. In *Proceedings EMLNLP-CONLL.*