Quantifying and Measuring Morphological Complexity

McWhorter (2001) contends in the lead article of an issue of *Linguistic Typology* that creole grammars are the world's simplest grammars. He proposes a metric of complexity and argues that creoles are simpler according to it, because as young languages they have not yet acquired the "ornament" of complicating distinctions (morphophonemic, syntactic, etc.) deposited on older languages by millennia of diachronic drift. In this work I suggest how we might attach some actual numbers to the issues in this debate, by offering a formal characterization of linguistic complexity as an improvement over McWhorter's relatively impressionistic approach.

I propose that we will be best served by employing *Kolmogorov complexity* (Solomonoff 1964, Kolmogorov 1965) as our metric, a quantifiable definition of descriptive complexity which is studied in algorithmic information theory. This notion, when applied to linguistic systems, accomplishes what McWhorter intends with his metric, but in a more precise and principled way. I indicate how, in the case of natural language morphology, we can compute concrete approximations of Kolmogorov complexity by using Goldsmith's (2001) *Linguistica* software. *Linguistica* is a tool which automatically infers the morphology of a language from a corpus of text, producing a lexicon of each form of every word, broken down into stems and affixes. In so doing, it gives us enough information to determine the proportion of a lexicon's complexity which is due to affixes and their distributions and alternations. This measure of morphological complexity is described by the following formula:

$$morphocomplexity = \frac{DL(affixes) + DL(signatures)}{DL(lexicon)},$$

where *DL(x)* is the "description length" of *x*, an approximation of Kolmogorov complexity computed by *Linguistica*, and where *signatures* refers to the descriptions of affix distributions that *Linguistica* infers.

I report some preliminary results in using this method to measure the morphological complexity of a sample of languages (so far, Latin, French, English, and Haitian Creole). The measurements obtained agree with the general consensus on the relative complexity of these languages' morphological systems. I also indicate some of the current limitations of this approach, and what future work is needed to address them. Lastly, I offer speculation on how we might approximate the Kolmogorov complexity of grammatical components other than morphology.

References

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27: 153-189.

Kolmogorov, A. N. 1965. Three approaches to the quantitative definition of information. *Problems in Information Transmission* 1(1): 1-7.

McWhorter, John. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 5: 125-66.

Solomonoff, R. J. 1964. A formal theory of inductive inference. *Information and Control* 7: 1-22, 224-54.