An information-theoretic approach to assess linguistic complexity

Benedikt Szmrecsanyi (FRIAS)

For a long time, most linguists considered all languages to be, on the whole, equally complex. This dogma has recently been challenged. Unfortunately, much of the recent controversy in regard to complexity (in)variance draws on empirical evidence that is subjective and/or fairly labor-intensive to obtain.

In this paper, I explore an idea that has been put forward by mathematicians and computer scientists (e.g. Juola 1998; Juola 2008) to use an unsupervised, algorithmic, information-theoretic measure for assessing linguistic complexity. I aim to evaluate this measure from the philologically responsible point of view of a trained linguist. The measure essentially boils down to the notion of KOLMOGOROV COMPLEXITY, which is defined as the length of the shortest algorithm that will exactly generate a given string or text; conveniently, most modern desktop computing compression tools draw on a variant of Kolmogorov complexity estimation to reduce file sizes. In the realm of linguistics, Kolmogorov complexity is all about linguistic SURFACE complexity, and thus I emphasize right at the outset that Kolmogorov complexity is completely agnostic about *deep* form-meaning relationships or such like. This agnosticism, however, is compensated by unparalleled algorithmical objectivity and scalability. The idea is to obtain numerical estimates of the relative informativeness and entropy (in technical parlance, SHANNON ENTROPY) of text samples, and to interpret higher observed entropy levels as being commensurate with higher levels of linguistic surface complexity.

To furnish a case study, I analyze a parallel text database that samples the Gospel according to Mark in a convenience sample covering six languages (Esperanto, Finnish, French, German, Hungarian, and Jamaican Patois) and some ten varieties of English, such as Wycliffe's Bible, the King James Version, or the Bible in Basic English. I subsequently marshal several measurements of Kolmogorov complexity in the database to assess

(i) overall surface complexity

(ii) surface complexity at the morphological tier

(iii) and surface complexity at the syntactic tier.

I demonstrate that Kolmogorov complexity measurements yield linguistically meaningful results, and provide complexity rankings that are in line with what more orthodox complexity notions would lead one to expect. I conclude by considering the advantages and drawbacks of the method, and by sketching directions for future research.

References

Juola, Patrick. 1998. "Measuring linguistic complexity: the morphological tier." *Journal of Quantitative Linguistics* 5 (3): 206-213.

Juola, Patrick. 2008. Assessing linguistic complexity. In *Language Complexity: Typology, Contact, Change*, ed. Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson. Amsterdam, Philadelphia: Benjamins.