Grammatical complexity of natural languages: The case of morphology

Witold Kieraś (University of Warsaw)

The aim of the paper is to present basic issues connected with the idea of measuring grammatical complexity of natural languages. The problem is important from different points of view: linguistic (both computational and theoretical), psychological, philosophical and for cognitive science in general. Solutions for this problem can lead us to answers about capabilities of human mind and it's linguistic competence.

At least since Hockett (Hockett, 1958) the thesis of the equivalent grammatical complexity of natural languages (understood as a ratio of morphological complexity and syntactic complexity) is widely approved. The natural argument for the thesis is that "all languages have about equally complex jobs to do, and what is not done morphologically has to be done syntactically" (Hockett, 1958). But recently linguists put the thesis in doubt (McWhorther, 2001). Different methods of estimating overall grammatical complexity (as well as morphological and syntactic complexity separately) have been applied. We focus on the information-theoretic approach to the problem of morphological complexity.

The information-theoretic approach to the problem was proposed by P. Juola (Juola, 1998). Juola's complexity metric uses so called morphological degradation: each token in the text sample is replaced by a random number between 1 and a number of types in the sample. The original and the degradated corpora are compressed using a popular compression tool. According to Juola, the morphological complexity metric for a given language is a ratio of the original compressed corpus to it's compressed degradated version. The assumption is that "by inflating the information content of the morphological tier, languages with a regular, informative, morphology will have their information content greatly increased relative to the information contained in the raw, unaltered sample" (Juola, 1998). So the ratio should be higher for languages of complex morphology.

The first goal of this paper is to evaluate the metric developed by Juola on two multilingual corpora focusing on European languages: the selection of European Union booklets translated into all UE languages and the Multext-East corpus of Orwell's "1984" translated into ten Central European languages (plus English original) (Erjavec, 2004). The second goal is to enhance the metric so it would work on wordforms rather than tokens. The motivation for such enhancement is that some languages (including Polish) have a very high level of homography (see Hajič 2000), which means that the Juola metric gives a lower complexity results for such languages than for those of low homography (even if the latter are less complex). Since Multext-East corpus has morphosyntactic annotation, the process of morphological degradation can be adopted to wordforms, which eliminates the homography problem and gives more precise results for morphological complexity.

Bibliography

Erjavec, T. (2004). Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Fourth International Conference on Language Resources and Evaluation, LREC'04*, Paris. ELRA.

Hajič, J. (2000). Morphological tagging: Data vs. dictionaries. In *Proceedings of the NAACL'00*, Seattle, WA.

Hockett, C. (1958). *A Course in Modern Linguistics*. Macmillan, New York.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5:206–213.

McWhorther, J. (2001). The world's simplest grammars are creole grammars. *Linguistic Typology*, 5:125–166.