Morphological analysis and corpus based lexicon development for Zulu

Sonja Bosch & Laurette Pretorius

University of South Africa – South Africa

The aim of this paper is to address aspects of an on-going project on the development of large-coverage grammatical and lexical resources for Zulu. This paper discusses the application of an existing rule-based, finite-state morphological analyser prototype ZulMorph in semi-automating the mining of the growing stock of electronic text corpora of the Zulu language for non-rule based behaviour. The work entails the exploitation of a guesser variant of the morphological analyser for the identification of various idiosyncrasies with regard to morphophonological rules. The necessity of the guesser variant becomes obvious in cases where the morphological analyser fails to analyse the word because the standard rules do not apply or do not exist. Examples are palatalisation rules for locative formation as well as verbal extension sequences. The semi-automated procedure makes provision for bootstrapping the morphological analyser to include newly extracted linguistic information from corpora. By means of human intervention the necessary information for these idiosyncrasies can be added to the machine-readable lexicon that plays a central role as lexical resource. It is shown how a machine-readable lexicon is in turn enhanced with the information acquired and extracted by means of such corpus analysis. The procedure is applied to a Zulu development corpus and the results are given and discussed.