The development of a rule-based lemmatiser for Setswana: The noun

Karien Brits (School of English, Adam Mickiewicz University, Poznań) & Rigardt Pretorius (School of Languages, North-West University)

Research in the field of Human Language Technology (HLT) is enjoying rapid growth the past few years with the support of the South African government (www.dac.gov.za/about_us/cd_nat_language/language_planning/hlt/english.htm). Part of this development is a rule-based lemmatiser for Setswana, one of the eleven official languages of South Africa. Setswana is in the privileged position of being extensively described in Krüger (2006) and Doke (1955), which was one of the main reasons to follow the rule-based approach in developing the lemmatiser. The first part of this paper focuses on the application of grammar in the programming; in the second part, the results of the lemmatiser (regarding nouns) are discussed.

Concerning the grammar, it is necessary to mention two important terms elements and hierarchy. The elements are the grammatical morphemes (prefixes and suffixes) and lexical morphemes (the root and stem) and the arrangement (hierarchy) of these elements are set Krüger (2006). A crucial step in developing a lemmatiser is defining the lemma in Setswana, in other words determining which elements should be included in the lemma and the second step is how they should be removed (and here the hierarchy plays an important role). In the first part of the paper, some examples of morphological analyses of nouns in Setswana will be given to explain the approach and the hierarchy. This hierarchy in the Setswana will be used to explain the process followed in the developing of the lemmatiser – thus the right sequence of removing the grammatical morphemes.

For this project, the noun lemma is defined as the simplest stem (i.e. the stem with the fewest inflections in the step before the root is identified) in the singular form without any augmentative, feminine, diminutive, or locative suffixes. The deverbative suffixes, however, will remain intact because the lemma should be in the same part-of-speech as the word to be lemmatised.

The Setswana grammar is described and tested thoroughly over the years, but it was the first time that it was put to such a test. In the last section, the results of the lemmatiser are discussed and reasons of errors, connected to unexpected sound changes and doubling of morphemes, will be explained.

Bibliography

Cole, D.T. 1955. *An introduction to Tswana grammar*. Johannesburg: Longman.
Krüger, C.J.H. 2006. Introduction to the morphology of Setswana. München: Lincom.