The structural-functional approach in developing a rule-based lemmatiser for Setswana

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. In the first part of the paper, the Setswana grammar and it application in the programming will be discussed briefly; in the second part, there will be looked at the results of the lemmatiser.

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 and which excluded 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 verbs 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 verb lemma is defined as the stem in the infinitive form, thus without any prefixes (i.e. negative, congruency, aspectual or temporal morphemes), and any suffixes such as the neuter-passive, iterative, causative, applicative, reciprocal, perfective, passive suffixes, or the terminatives *-e* and *-ng*. The only productive suffixes included in the lemma are the unmarked verbal suffix *-a*. The unproductive suffixes and some of the semi-productive suffixes, i.e. the denominative and reversive suffixes will be left intact, as these suffixes are no longer very productive in Setswana.

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 change in the hierarchy and unexpected 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.