Noun and verb guessing for automatic POS tagging in Sepedi

DJ Prinsloo & Gertrud Faasz (University of Pretoria)

One of the main challenges to POS tagging in Sepedi lies in the automatic identification of verb stems and nouns. Our aim is to devise a methodology that can maximally utilize their morphosyntactic features in order to automatically identify and label them as such in any given text.

The presentation will be introduced by a brief overview of the nominal and verbal linguistic patterns of Sepedi. This will be followed by an outline of the specific problems of annotating Sepedi texts morphosyntactically, caused by the fact that previously defined concepts of POS cannot just be 'copied', but must be adapted to the disjunctive orthography and the noun class system of Sepedi. We use both morphologic and syntactic labels to enable us to use token-based taggers like the Treetagger (Schmid 1994) and our tagset contains information on noun classes. We will demonstrate a hybrid approach (cf. Heid and Prinsloo 2005), aiming at partially rule-based, partially statistical identification and disambiguation procedures. The rule-based component covers the scarce environments, while the statistical component caters for the frequent cases. Because of the noun class system, especially nouns and verbs appear in rather scarce environments, hence their identification is done rule-based. Utilizing the current procedures (not all training data planned is yet available and not all rule-based components are fully-developed), the overall tagging process already reaches about 92% precision.

Our work will be compared with that of De Schryver and De Pauw (2007) and with the tagset designed by Van Rooy and Pretorius (2003).

Our approach for Sepedi *verb guessing* consists of

- Reversal of morphophonological processes caused by affixation
- Detection of verb roots by identification of typical verbal suffixes and combinations of such suffixes (of which up to 400 forms may be used in Sepedi)
- Attachment of frequent suffixes to the identified roots and consequent look-up in a lexicon and in the *University of Pretoria Sepedi Corpus* (PSC), thereby generate basic verb stem forms.

For *noun guessing* we use three criteria i.e. class prefixes, nominal suffixes and syntactic environment as basic pillars. The procedure consists of

- Utilization of the specific grammatical relation between noun classes, i.e. singularity versus plurality.
- Consideration of the three nominal suffixes *locative*, *augmentative/feminine* and *diminutive*, and rules for the formation of nouns that are derived from verbs.
- Consideration of the co-text of the noun candidate

Both guessers are processed in combination to allow the guessing process to suggest candidates as belonging to both word classes.

The discussion will be concluded by an evaluation of the automatic guessing of nouns and verb stems and of the overall tagging process.

Bibliography

De Schryver, G-M and De Pauw, Guy. 2007. Dictionary Writing System (DWS) + Corpus Query Package (CQP): The Case of *Tshwanelex*. *Lexikos* 17 226-246

Prinsloo, D. J. and U. Heid. 2005. Creating Word Class tagged Corpora for Northern Sotho by Linguistically Informed Bootstrapping. *Conference for Lesser Used Languages and Computer Linguistics, EURAC research, European Academy*. Bolzano, Italy. 27th October-28th October 2005.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees, in: *Proceedings of the International Conference on New Methods in Language Processing* Manchester, UK. pp. 44 – 49.

Van Rooy, B. and R. Pretorius A. 2003. A word-class tagset for Setswana. *Southern African Linguistics and Applied Language Studies 2003, 21 (4): 203 – 222.*