The need for a large(r) Afrikaans treebank

Peter Dirix
Centre for Computational Linguistics, KU Leuven, Belgium
E-mail: peter.dirix@kuleuven.be

Linguistics nowadays relies more and more on the existence of large corpora, and on the tools to search them. Afrikaans is still a low-resource language in this respect, although the situation has improved considerably since the launch of the VivA corpus portal\(^1\) in 2016. Currently it contains a lemmatized and morphosyntactically tagged corpus of about 300 million tokens. However, the corpus has been automatically tagged, and the accuracy of lemmatization is only 90\%, whereas the morphosyntactic annotation about 75\%, as it has been trained on a very small manually verified annotated corpus. This often leads to a high number of false positive hits. Moreover, the automated tools needed to do the annotation are not freely available for people who want to tag their own data.

At this point, some free tools are available at the github repository AfriTools\(^2\). This repository currently consists of (i) a manually checked lexicon of about 270 K tokens with their lemma and morphosyntactic properties using a slightly modified version of the tag set defined in Pilon (2005), (ii) a rule-based tokenizer based on the Dutch one originally developed for the METIS-II project (Dirix et al. 2005), and (iii) a parameter file for Afrikaans to be used with the TreeTagger lemmatizer and morphosyntactic tagger (Schmid 1994). The model was trained on an automatically tagged version of the Taalkommissie corpus, limiting the lemma/morphosyntactic assignments for lexicalized words to those listed in the accompanying lexicon. On a test set of about 3,200 words from Wikipedia, the lemmatizer reached an accuracy of 99.7\%, while the morphosyntactic assignments were 94.9\% correct.

While these are certainly useful resources, there are two problems. First, there is no large corpus which also encodes the syntactic structure. The only attempt to do this was the AfriBooms treebank (Augustinus et al. 2016), which was later re-released as part of the Universal Dependencies project (Dirix et al. 2017). However, with only 45 K tokens of government text, it is both too small and representative of too narrow a domain. When studying infinitivus pro participio (IPP), for instance, only two examples were found. Looking at, for example, insubordination, complex vs. simplex initials, or word order in verb clusters, researchers run into the issue of the non-existence of a large and diverse treebank for Afrikaans and have to classify a large number of sentences manually.

\(^{1}\) https://www.viva-afrikaans.org
\(^{2}\) https://github.com/peedirix/AfriTools
In order to deal with this issue, I would like to plead for the creation of a large treebank for Afrikaans, modelled, for example, on the Spoken Dutch Corpus (CGN) (Schuurman, 2003), which contains about 10 million tokens (of which about 1 million are manually annotated). The corpus should be balanced between written and spoken Afrikaans, and also cover all main variants, including Namibian and Cape Afrikaans. The syntactic annotation should ideally happen in the Universal Dependencies framework (De Marneffe et al. 2021), as this will allow it to profit from this framework’s substantial tool development package usable for all languages. The result will also include one or more parsers which will allow people to annotate their own data, and which can hopefully also better deal with non-standard varieties of Afrikaans.

References


