Stellenbosch Papers in Linguistics Plus, Vol. 67, 2023, 1-2
doi: 10.5842/67-1-1001

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

© 2023 The author. This work is licensed under the Creative Commons Attribution 4.0 International License.

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

Augustinus, Liesbeth, Peter Dirix, Daniel van Niekerk, Ineke Schuurman, Vincent Vandeghinste, Frank Van Eynde & Gerhard van Huyssteen. 2016. AfriBooms: An Online Treebank for Afrikaans. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk & Stelios Piperidis (eds.). *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož: European Language Resources Association. pp. 677-682.

De Marneffe, Marie-Catherine, Christopher Manning, Joakim Nivre, & Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2): 255-308. https://doi.org/10.1162/coli_a_00402

Dirix, Peter, Liesbeth Augustinus, Daniel van Niekerk & Frank Van Eynde. 2017. Universal Dependencies for Afrikaans. In Marie-Catherine de Marneffe, Joakim Nivre & Sebastian Schuster (eds.). *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*. Gothenburg: Association for Computational Linguistics. pp. 38-47.

Dirix, Peter, Vincent Vandeghinste & Ineke Schuurman. 2005. METIS-II: Example-based machine translation using monolingual corpora – System description. In Michael Carl & Andy Way (eds.). *Proceedings of MT Summit X, Workshop on Example-Based Machine Translation*. Phuket. pp. 43-50.

Pilon, Suléne. 2005. Outomatiese Afrikaanse woordsoortetikettering. Master's thesis, North-West University, Potchefstroom.

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

Schuurman, Ineke, Machteld Schouppe, Heleen Hoekstra & Ton van der Wouden. 2003. CGN, an annotated corpus of spoken Dutch. In Anne Abeillé, Silvia Hansen-Schirra & Hans Uszkoreit. *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*. Budapest: Association for Computational Linguistics. pp. 101-108.