The TypeCraft (TC) Akan Corpus

Dorothee Beermann

Release 1.0

1. License and Legal Issues

This corpus is distributed solely for non-commercial, non-profit educational and research use. Release 1.0 consists of a derivative compilation of multiple annotated texts created by linguistic graduate students as part of their class work. The original texts were created between 2007 – 2013 at the Department of Linguistics, NTNU, Trondheim, Norway.

2. How to get the TypeCraft Akan Corpus

The release represents a subcorpus of the TC Akan corpus which has been created by the TypeCraft project. The project maintains TypeCraft - The Interlinear Glossed Text Repository (https://typecraft.org) which is a linguistic online service. Release 1.0 can be downloaded from:
https://typecraft.org/tc2wiki/The_TypeCraft_Akan_Corpus

It is an XML data set. The TypeCraft xsd-schema you find at:
https://typecraft.org/typecraft.xsd

Poio-api (https://github.com/cidles/poio-api) converts file formats like TypeCraft XML, Elan's EAF, Toolbox files, and others into annotation graphs as defined in ISO 24612. Poio API is a free and open source Python library to access and search data from language documentation in your linguistic analysis workflow.

The TypeCraft Importer allows you to import the corpus into the TypeCraft Editor for further annotation, or for export to other formats.

3. Description of the TypeCraft Akan Corpus

The Release 1.0 of the TC Akan Corpus consists of 41 short texts, mostly linguistic sentence collections, corresponding to 669 sentences. Two of the released texts are transcribed recordings of students narrating a video. The students doing the original work were native-speakers of Akan. The material was curated, starting in 2016 over the period of 1 ½ years. For the curation expert linguists worked together with student annotators, native and non-native speakers, to achieve a better consistency of the original data. We will say more about the type and depth of annotation below. On a 4 point scale from green (high quality), yellow, orange and red (should not be used for research), we would like to characterise the Release 1.0 as a yellow corpus by which we mean that it can be used for research with some care.

4. Data structure

All texts carry morpheme-based annotations as well as POS tags. The original texts have been translated to English.The whole corpus had been privacy masked and meta data is provided.

For the annotation of the TC Akan corpora we use an annotation set consisting of 60 POS tags and 123 gloss tags. Chart 1 shows the most frequently assigned POS tags and Chart 2 shows the corresponding for Gloss tags.
The TC POS and GLOSS tag sets can be found here:
https://typecraft.org/tc2wiki/Special:TypeCraft/POSTags/ and
https://typecraft.org/tc2wiki/Special:TypeCraft/GlossTags/

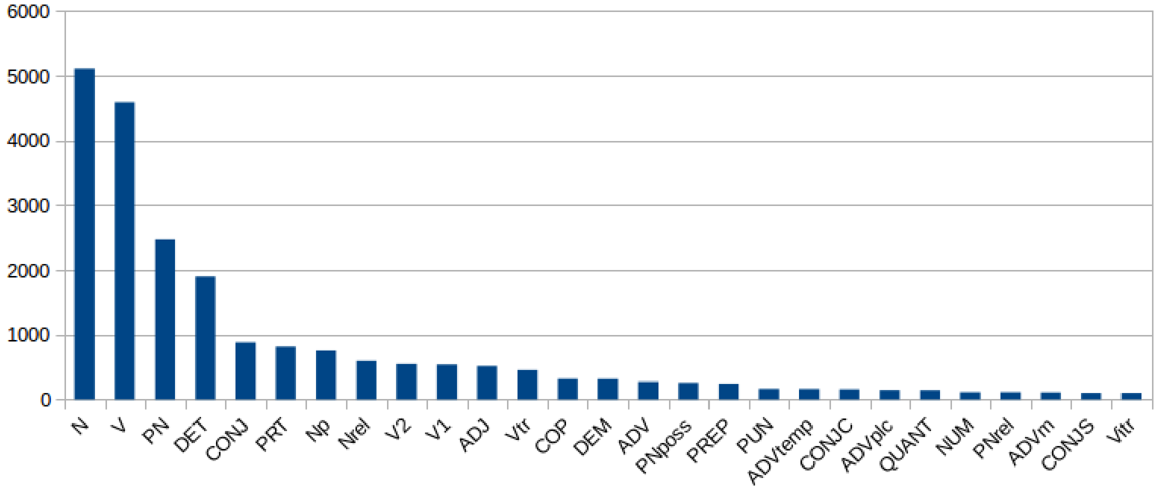

Chart 1 TC Akan corpus - The most frequently assigned POS tags (>100)

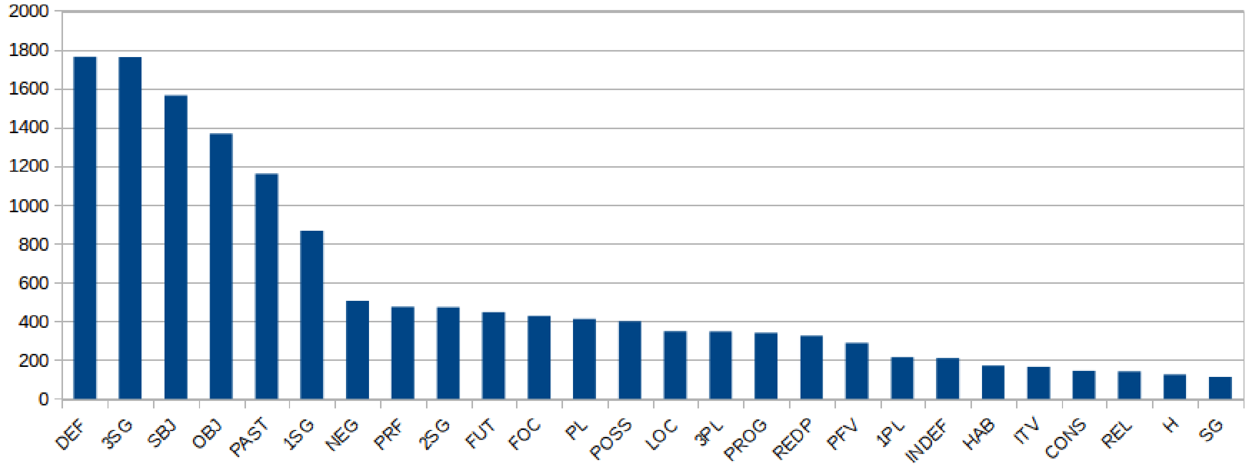

Chart 2 TC Akan corpus - The most frequently assigned Gloss tags (>100)

5. Authors, Citation and Contact Information

The TC Akan corpus was created by Dorothee Beermann. Special thanks for their support goes to Associate Professor James Essegbey, University of Florida in Gainsville and The Ghanaian Student Association at the Norwegian University of Science and Technology.

The corpus should be cited as follows:

Dorothee Beermann (2018). The TypeCraft Akan corpus, Release 1.0.
TypeCraft – The Interlinear Text Repository.
https://typecraft.org/tc2wiki/The_TypeCraft_Akan_Corpus

Please address all questions, comments and suggestions to
Dorothee Beermann (dorothee.beermann@ntnu.no)