I. The TypeCraft Akan corpus

We start with an example:

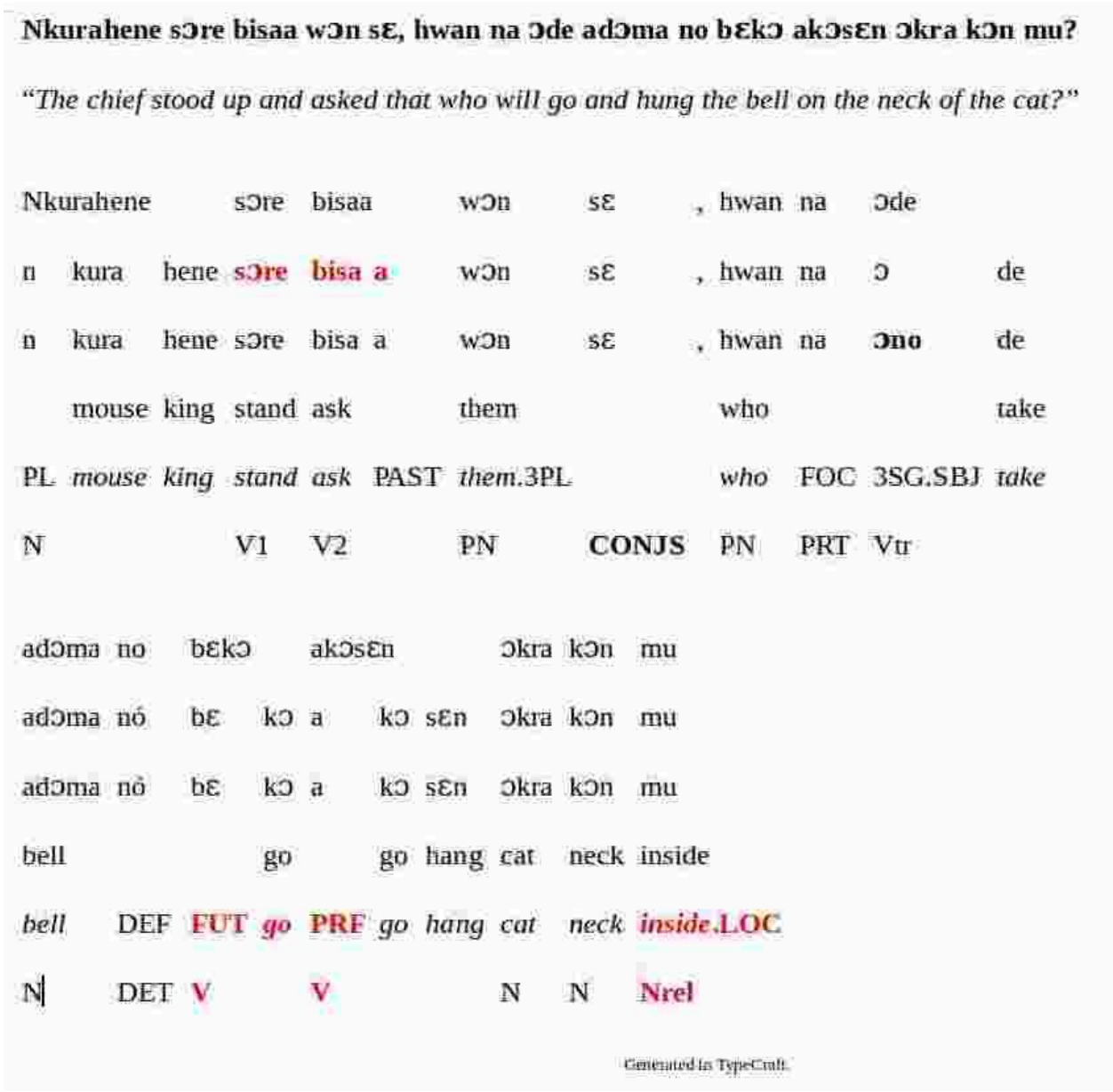


Figure 1 An Akan (ISO- 639-3 'aka') sentence example as it appears in the TC Editor

The TC Akan corpus consists of 80893 words of which 9347 (11.55%) received a POS tag. Most of these 9000 words are also glossed. The annotators were all linguistic graduate students at NTNU, and all of them were speakers of one of the dialects of Akan: Akuapem, Fante ('fant'), Twi ('twi) and Abron (abr) (a difference not marked in the

corpus). The corpus consists of transcribed oral narratives or radio programs, and transcriptions of Akan movies, as well as of linguistic sentence collections. Especially in the latter category punctuation does not play a role; which means it is either absent, or when present it does not receive an annotation. This explains the low frequency of the PUN tag in Table 1.

The POS tags of our Akan corpus are distributed as shown in Table 1

Table 1 Akan POS tags assigned more than a 100 times

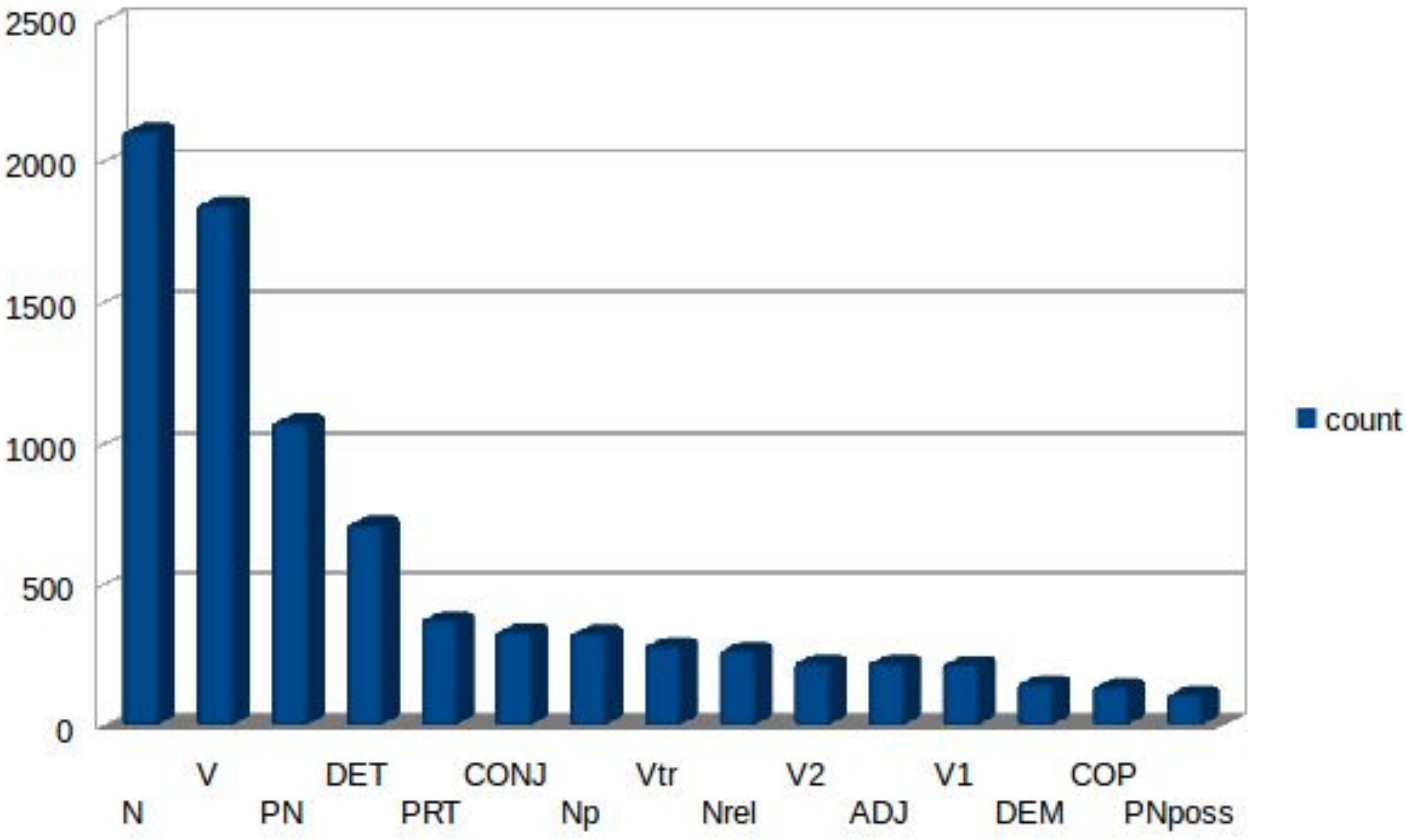
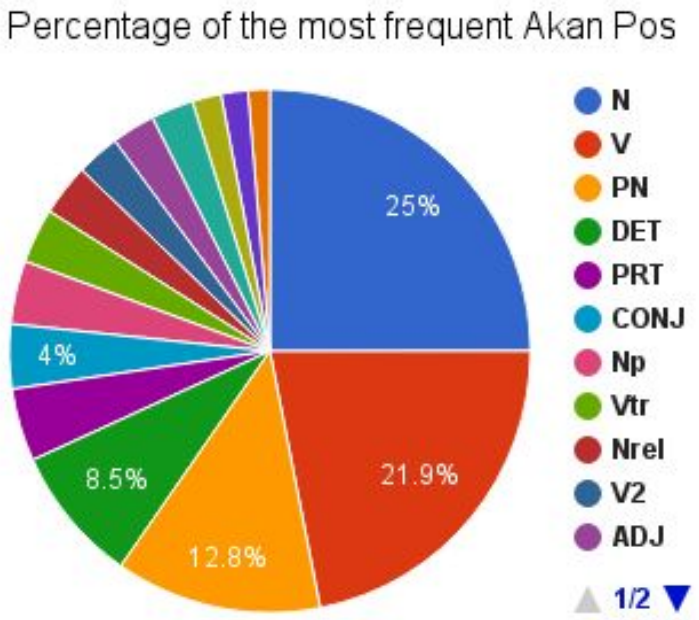


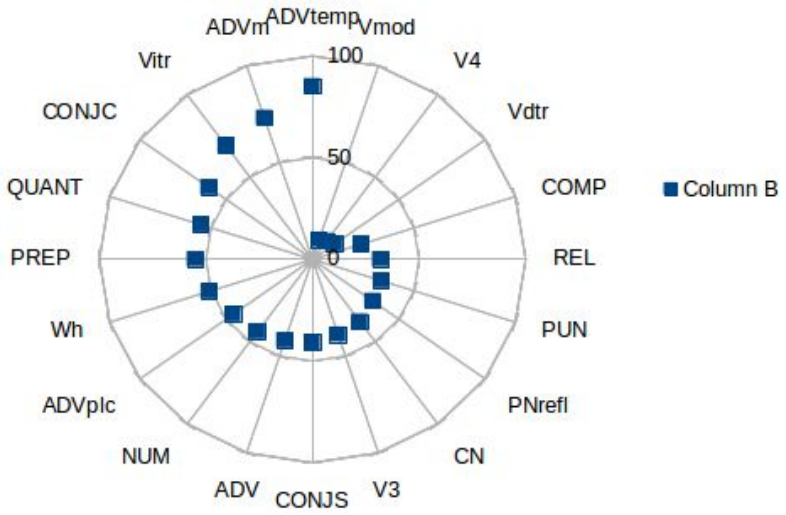
Table 2 Percentages for the most frequent Akan POS tags



If we can make any predictions on the basis of the distribution of 9000 pos tags this presumably means that the probability of any given word in Akan to be a noun is 0.25 while the probability of it being an adjective is 0.04. More impressionistically speaking, judging from the present quality of our corpus, the likelihood that something that has been annotated as a noun is in fact a noun and not a verb is high, as is the likelihood of something annotated as a pronoun not being a noun or a verb, However, the likelihood of something that has been annotated as a determiner and not as a pronoun to be in fact a determiner is considerably lower (same orthographic form).

As for the less frequent POS tags (Table 3) not much can be said at this point. But notice for example that the annotators have tried to distinguish between subordinating and coordinating conjunctions. Some also tried to distinguish sentential complements from other embedded sentence types. So there is the attempt of some linguistic depth to the corpus.

Table 3 Distribution of Akan low frequency POS tags



We now look at the gloss tags.

Table 4 Overall frequency of the Akan gloss tags

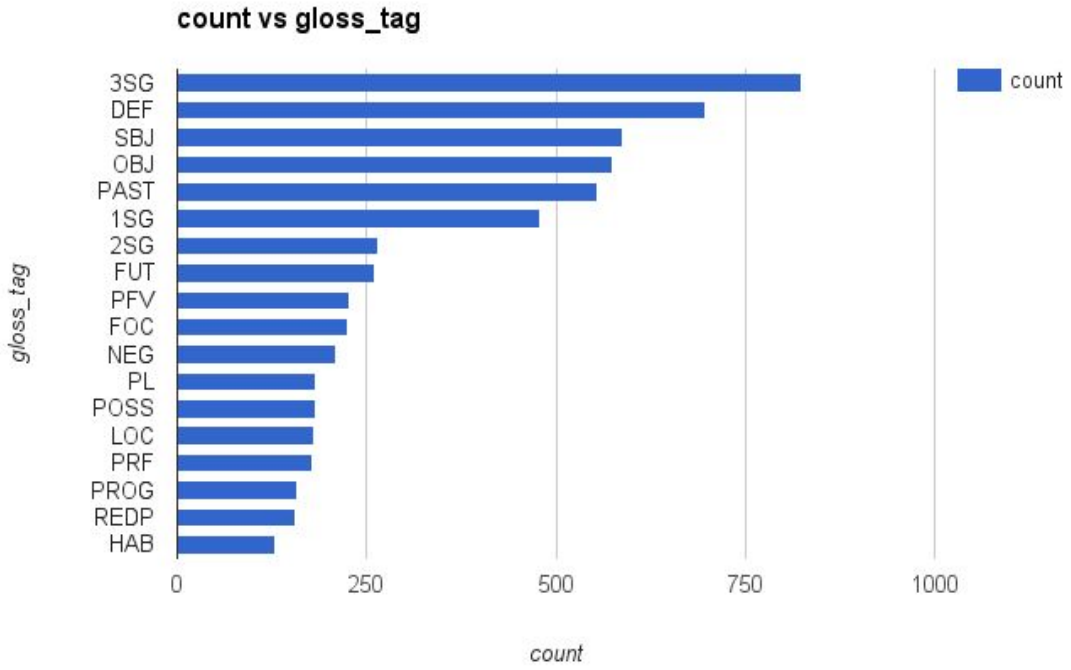


Table 5 Distribution of gloss tags relative to POS tags

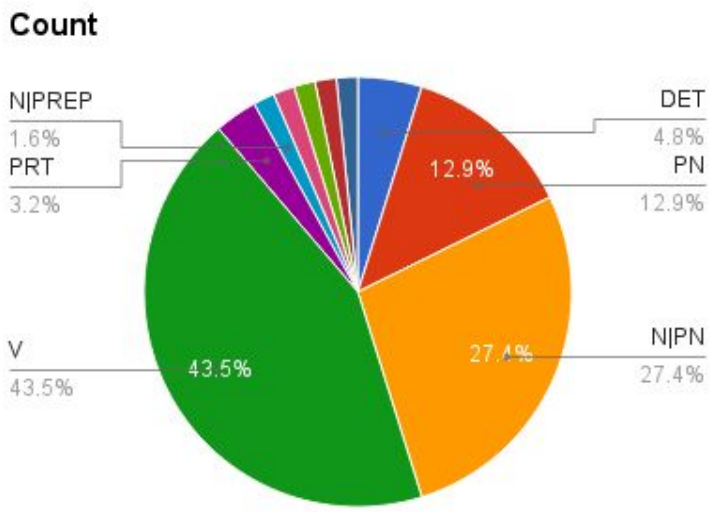


Table 6 Individual distribution of the most frequent Akan Gloss tags