Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 294–297
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Developing a Twi (Asante) Dictionary from Akan Interlinear Glossed Texts

Dorothee Beermann, Lars Hellan, Pavel Mihaylov, Anna Struck
Department of Language and Literature, Norwegian University of Science and Technology, Norway
{lars.hellan, dorothee.beermann}@ntnu.no

Abstract

Traditionally, a lexicographer identifies the lexical items to be added to a dictionary. Here we present a corpus-based approach to dictionary compilation and describe a procedure that derives a Twi dictionary from a TypeCraft corpus of Interlinear Glossed Texts. We first extracted a list of unique words. We excluded words belonging to different dialects of Akan (mostly Fante and Abron). We corrected misspellings and distinguished English loan words to be integrated in our dictionary from instances of code switching. Next to the dictionary itself, one other resource arising from our work is a lexicographical model for Akan which represents the lexical resource itself, and the extended morphological and word class inventories that provide information to be aggregated. We also represent external resources such as the corpus that serves as the source and word level audio files. The Twi dictionary consists at present of 1367 words; it will be available online and from an open mobile app.

Keywords: lexicographical resources, Akan data, Akan lexicographical model, online dictionary

1. Introduction

The availability of lexicographical data is essential for the development of digital applications for less resourced languages. Akan, although a well described language in linguistic terms and a language not without important lexical resources, is still in terms of digital means a lesser resourced language. Here we describe the compilation of a small digital dictionary for Twi (Asante) based on an interlinear glossed corpus of Akan. This being a goal in itself, we at the same time hope to strengthen the development of Interlinear Glossed Text corpora for lesser resourced languages as a resource for the digital deployment of lexicographical data.

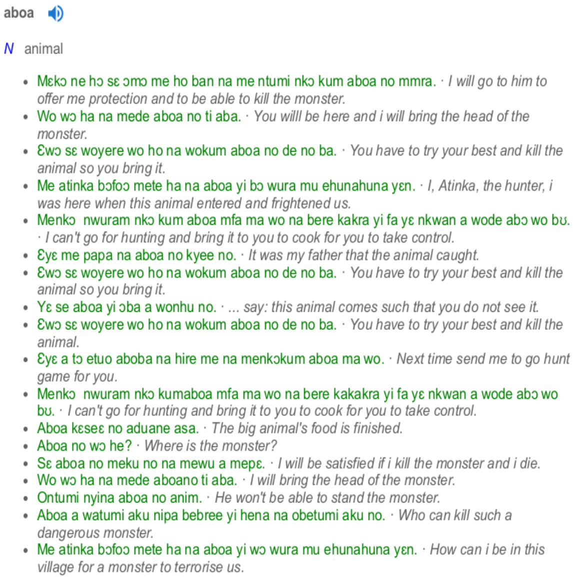
Akan is a Kwa language and one of the official languages of Ghana. Foremost among its dictionaries is Christaller (1881) which was the first dictionary of Twi, and still is by far the most comprehensive one. Less comprehensive is Anyidoho et al. 2005. Neither of these two dictionaries is in a format easily amenable to the development of further digital resources.

Online dictionaries for some Ghanaian languages are available from the Ghana Institute of Linguistics Literacy and Bible Translation (GILLBT)[1], including one for Ghanaian Kusaal with over 4000 entries. Its entries may provide information about the part of speech and offer an English translation of the head word. They may give an example in the source language, as exemplified in (1) and (2) below, but entries may also consist just of the head word.

(1) pa'al-1 v teach, *Tinam pa'an biis nɛ yin yɛlsieba amaa asɛɛ pa'annib pa'al ban na niŋ si'em sɔbi li*

(2) pa'al-2 *v* show, *Msaam mɔrimini keŋ pa'al kuob la zɛm si'em*

The present approach describes the compilation of a Twi (Asante) online dictionary from a TypeCraft corpus of Interlinear Glossed Texts. A view of an entry analogous to the above is shown in Figure 1 and will be described in the main body of the article in more detail.



aboa

N animal

- Mɛkɔ ne hɔ sɛ ɔmɔ me ho ban na me ntumi nkɔ kum aboa no mmra. · *I will go to him to offer me protection and to be able to kill the monster.*
- Wo wɔ ha na mede aboa no ti aba. · *You willl be here and i will bring the head of the monster.*
- Ɛwɔ sɛ woyere wo ho na wokum aboa no de no ba. · *You have to try your best and kill the animal so you bring it.*
- Me atinka bɔfoɔ mete ha na aboa yi bɔ wura mu ehunahuna yɛn. · *I, Atinka, the hunter, i was here when this animal entered and frightened us.*
- Menkɔ nwuram nkɔ kum aboa mfa ma wo na bere kakra yi fa yɛ nkwan a wode abɔ wo bʊ. · *I can't go for hunting and bring it to you to cook for you to take control.*
- Ɛyɛ me papa na aboa no kyee no. · *It was my father that the animal caught.*
- Ɛwɔ sɛ woyere wo ho na wokum aboa no de no ba. · *You have to try your best and kill the animal so you bring it.*
- Yɛ se aboa yi ɔba a wonhu no. · *... say: this animal comes such that you do not see it.*
- Ɛwɔ sɛ woyere wo ho na wokum aboa no de no ba. · *You have to try your best and kill the animal.*
- Ɛyɛ a tɔ etuo aboba na hire me na menkɔkum aboa ma wo. · *Next time send me to go hunt game for you.*
- Menkɔ nwuram nkɔ kumaboa mfa ma wo na bere kakakra yi fa yɛ nkwan a wode abɔ wo bʊ. · *I can't go for hunting and bring it to you to cook for you to take control.*
- Aboa kɛseɛ no aduane asa. · *The big animal's food is finished.*
- Aboa no wɔ he? · *Where is the monster?*
- Sɛ aboa no meku no na mewu a mepɛ. · *I will be satisfied if i kill the monster and i die.*
- Wo wɔ ha na mede aboano ti aba. · *I will bring the head of the monster.*
- Ontumi nyina aboa no anim. · *He won't be able to stand the monster.*
- Aboa a watumi aku nipa bebree yi hena na obetumi aku no. · *Who can kill such a dangerous monster.*
- Me atinka bɔfoɔ mete ha na aboa yi wɔ wura mu ehunahuna yɛn. · *How can i be in this village for a monster to terrorise us.*

Figure 1: The word "aboa", in the Twi (Asante) dictionary

Also to be described in the following is the compilation of the dictionary. Our corpus-based approach requires a routine of data cleaning and normalization: we extracted a list of unique words, excluded words belonging to different dialects of Akan (mostly Fante and Abron), corrected misspellings and distinguished English loan words to be integrated in our dictionary from instances of code switching, which is frequently found in spoken Akan also reflected in our corpus.

The paper is structured as follows: Section 2 describes data acquisition and the essential properties of the Akan corpus on which the dictionary is based, together with the basic lexicographical features to be represented in the dictionary. Section 3 presents a lexicographical Language Model for

[1] GILLBET's dictionaries like many other dictionaries can be accessed from the Webonary – Dictionaries and Grammars of the World : https://www.webonary.org/

Akan. Section 4 gives a brief evaluation of our present resources and describes the next phase of the development.

2. The Akan Source Data

2.1 The Corpus

The data used for this dictionary is based on a collection of Akan texts which are stored in TypeCraft - a user-driven online database for Interlinear Glossed Texts. Akan is one of the official languages of Ghana belonging to the Niger-Congo languages of West Africa, Twi (Asante) is one of the main dialects of Akan.

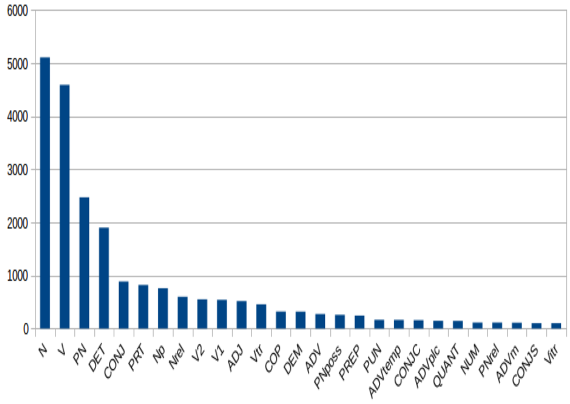
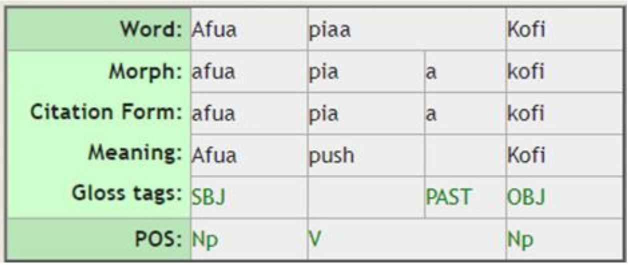


Figure 2 : TC-Akan corpus - Part of Speech labels

The corpus currently consists of 261 texts, narratives, transcribed dialogue, or linguistic collections, corresponding to 98 000 words. Most of these texts have been annotated for part of speech and are glossed. An entry in the TypeCraft database consists of a sentence, its translation, and several layers of annotation, see Figure 3 below. The most common part of speech tags used for Akan are shown in Figure 2 above.

The largest word class in our corpus are nouns, followed by single main verbs. We distinguish verbs as part of an SVC (V1, V2, V3) from single verbs and copulas. Texts which are fully curated will not contain items labeled as prepositions, which, as Figure 2 shows, still occurs as a label in non-curated parts of the corpus, although it is well-known fact that Akan makes only use of postpositions (relational nouns). We will discuss the morpho/syntactic labeling further in section 4.



Word:	Afua	piaa		Kofi
Morph:	afua	pia	a	kofi
Citation Form:	afua	pia	a	kofi
Meaning:	Afua	push		Kofi
Gloss tags:	SBJ		PAST	OBJ
POS:	Np	V		Np

Figure 3 : Example of an interlinear glossed sentence

2.2 The Lexical Entry

In its present form the entries of our lexicon contain the kind of information instantiated in (3):

```
(3)
[{
"word": "ankaa",
"pos": "N",
"translation": "orange",
 "audio": [ "F_9_anka", "M_9_anka" ],
 "examples": [ {
"original": "Kofi tɔɔ ankaa firii ankaa-
sensenetɔnfoɔ no hɔ.",
 "translation": "Kofi bought an orange from
the peeled orange sellers."
},
...
} ],
"id": 132,
"key": "ankaa"
```

In (3), next to the lemma itself we indicate the part of speech and provide an English translation. Most entries are accompanied by a word level audio file (if available we provide both a female voice and a male voice). Most entries come with a list of examples from the TypeCraft database, as shown in Figure 1.

2.3 Creating the Entries

Akan is a tone language. When creating the audio files for the dictionary, speakers were given sentences to read. In a second round we recorded word level data (for more information, see Van Dommelen and Beermann, 2019). Words written in isolation can correspond to several meanings depending on the lexical tone, but also dependent on the context they appear in. As part of the dictionary compilation, we added new words to the dictionary when an orthographic form corresponded to different tonal patterns and meanings.

The phonetic project that had allowed us to record speakers in phonetic experiments, ended while we still were working on the dictionary compilation. When we finally could revisit the data, some of the links between audio files, words, and the corpus were hard to reconstruct. We therefore used the word's orthographic form, as well as part of speech information and the word's meaning (rendered in English), to extract example sentences from the corpus. The corpus itself does not contain tone marking, which means that not in all cases words and their example sentences may have been aligned correctly.

3. The Akan Language Model

For the representation of the structure and external resources of the present dictionary, and in order to show the further development that the Twi (Asante) dictionary will take, we present a UML model[2] as shown in Figure 4. The only other African language for which we found a lexicographical language model is Xhosa (Bosch et al. 2018), a Bantu language spoken in South Africa. It is agglutinating in nature, while Akan is a Kwa language and belongs to the Atlantic branch of the Niger-Kongo languages. Like other Kwa languages, Akan does not make use of noun classes and derivational suffixes which is characteristic for Bantu languages. Kwa languages are known for their Serial Verb Constructions (SVC) where several verbs share a subject.[3] When part of an SVC, verbs have an inflection which is sensitive to their position in the cluster as well as to the SVC's temporal-aspectual features. This and other differences between these two language families carry over to their lexicographical models.

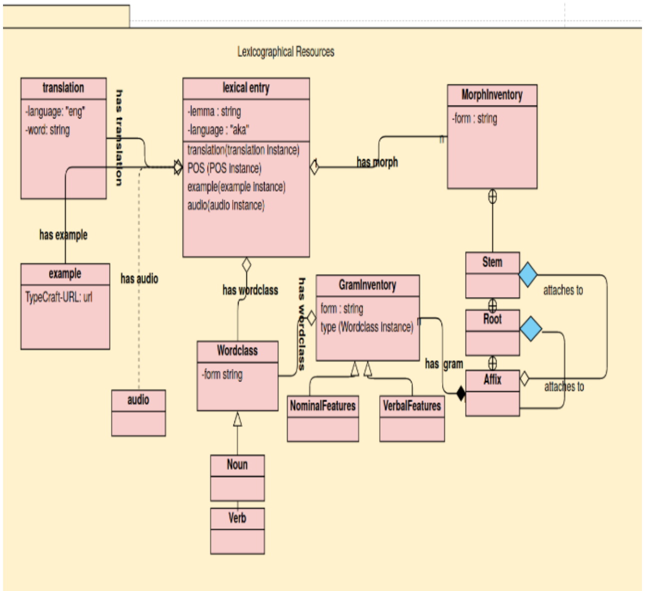


Figure 4: Akan lexicographical model

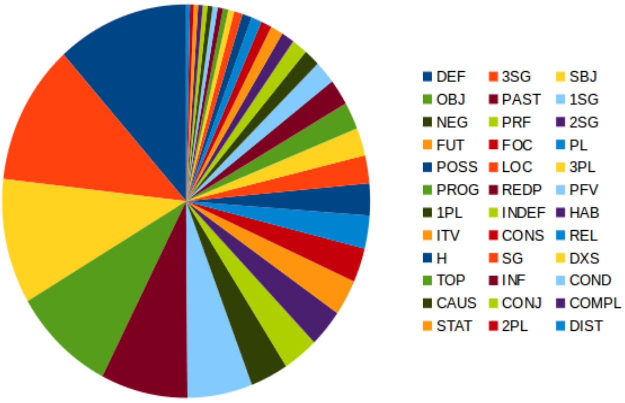


Figure 5: Morpho-functional annotations - TC Akan corpus

Given the present form of the dictionary, only a part of the grammatical information we have stored in our corpus is reflected in the dictionary. On the other hand, additional external information such as audio and corpus resources in the form of examples have been added to make the dictionary more user friendly.

Morphological information, although available in the corpus, has not yet been included in the dictionary. An overview over the grammatical features annotated in the corpus is given Figure 5. [4]

Interesting in Figure 5 is in this context not so much how frequently a certain label is used as the nature of the information that has been labeled. Especially for languages where a grammar book is not a standard commodity, the inclusion of morphological information in a dictionary is of special interest. The corpus will allow us to use this information as part of the Morph-and-Gram-Inventory which is part of our Akan lexicographical model as shown in Figure 4. Lemmas are shown to be related to the inflectional paradigms they are members of, and this information is stored relative to word class specification.

The *Morph Inventory* as well as the *Gram Inventory* can be populated by morpheme forms and the grammatical features they embody.

Which additional features are available is shown in Figure 5. We find form function pairs for most of the language's Tense-Aspect features, and for some of the grammatical relations. but still very little is known. Although we still lack information about the nominal morphology in our corpus, this information itself is readily available, and can be added to the corpus.

4. Evaluation and Outlook

The Twi (Asante) dictionary is a small resource which needs to grow in order to be a real resource. At present the dictionary contains word class information, in most cases an audio representation which gives an indication of the tonal properties of a word in its base form. In many cases we provide an extensive list of examples which are glossed and translated into English, which goes beyond what most Akan online dictionaries provide. Already in its present form the dictionary will be useful; especially if new words are added on a regular basis. The resource will be available as an online dictionary and through an open mobile dictionary app (Eckart et al., 2020).

Due to the in-depth annotated TC-Akan corpus we will be able to further extend the dictionary with morphological information. Akan verbs belong to different classes according to which inflectional paradigms unfold (for more information see Dolphyne 1988). We further observe that, when part of an SVC, verbs may be associated with additional inflectional patterns specific to their position in the verb cluster and the nature of the cluster itself. This means that we need to assign Akan verbs to several

[2] UML stands for *Unified Modeling Language*

[3] See Beermann and Hellan 2018

[4] SBJ=subject, 3SG=3Person, singular, DEF= definite, 1SG =1Person,singular, PAST=past tense, OBJ=object. A full list of the glosses can be found at TypeCraft: https://typecraft.org/tc2wiki/Special:TypeCraft/GlossTags

inflection paradigms dependent of whether they occur as a single verb or as part of an SVC. Using annotation mining, part of this information may be directly acquired from the Interlinear Glossed Text corpus. In addition, we will make use of the Akan data from the Leipzig Corpus Collection.[5] An LCC corpus contains randomly selected sentences which in the case of the 2018 Akan corpus come from the Wikipedia. Although our corpus resources are small, valuable information can be extracted. The LCC offers a basic language statistic for the corpus, and together with information from word sketches, to show collocational behavior, we hope to arrive at a more accurate and detailed analysis of our lexical data. This then represents the next step in the development of a Twi (Asante) corpus and the Twi dictionary.

5. Bibliographical References

Anyidoho, Akosua, et al. 2006. *Akan Dictionary*. Pilot project. University of Ghana.

Beermann, Dorothee. 2014. Data management and analysis for less documented languages. In Jones, Marion, and Connolly, C. (eds) *Language Documentation and New Technology*. Cambridge University Press.

Beermann, Dorothee, and Mihaylov, Pavel. 2014. Collaborative databasing and Resource sharing for Linguists. In: *Languages Resources and Evaluation*. Springer. 48. Dordrecht: Springer, 1-23.

Beermann, Dorothee, and Lars Hellan. 2018. West African Serial verb constructions: the case of Akan and Ga. In: Agwuele, Augustine, and Adams Bodomo (eds) *The Routledge Handbook of African Linguistics*. London and New York: Routledge. Pg. 207-221.

Bosch, S., Eckart, T., Klimek, T., Goldhahn, D., and Quasthoff, U. (2018). Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation* (LREC 2018), Miyazaki, Japan.

Christaller, J.G. 1881 (latest edition of 2013). *Dictionary of the Asante and Fante Language.* Basel: Basel Evangelical Missionary Society

Dolphyne, Florence A. 1988. The Akan (Twi-Fante) Language. Accra: Ghana Universities Press.

Eckart, Thomas, Sonja Bosch, Uwe Quasthoff, Erik Körner, Simon Kaleschke, Dirk Goldhahn. Usability and Acessibility of Bantu Language Dictionaries in the Digital Age: Mobile Access in an Open Environment. *Proceedings of LREC 2020, Marseille.*

van Dommelen, Wim A.; Beermann, Dorothee. (2019) Tonal properties of the Akan particle 'na'. *Proceedings of the 19th International Congress of Phonetic Sciences*.

6. Language Resource References

TypeCraft Akan corpus, Release 1.0: https://www.researchgate.net/publication/323998547_TypeCraft_Akan_Corpus_Release_1.0

Further TypeCraft Akan corpora, non-curated: https://typecraft.org/tc2wiki/Special:TypeCraft/PortalOfLanguages

[5] https://wortschatz.uni-leipzig.de/en/download/