Quantitative and Qualitative Analysis in the work with West African Languages

Dorothee Beermann **Lars Hellan**

NTNU, Trondheim, Norway

dorothee.beermann@ntnu.no lars.hellan@ntnu.no

Abstract

The work we report on is part of a larger research project which searches to combine generally accessible resources of African languages into common repositories and platforms on which property extraction from these resources can lead to new views and new insights into the phenomena addressed. Among these resources are the TypeCraft Interlinear Glossed Text Repository (TC) (Beermann & Mikailov 2014), the African languages corpora and search environment of the Leipzig Corpora Collection (LCC) (Goldhahn 2012), and resources from the multilingual verb valence project (Hellan et al. 2014). Dealing with these resources in a common digital infrastructure facilitates various types of linguistic processing and corpus methodologies for lesser-resourced languages. In the present paper we focus on semi-automatic acquisition of linguistic resources at part-of-speech, morphology and valency levels. In this way we support the aim of our project of increasing the access to data from African languages by providing accessibility to data at different levels of analysis that can inform linguistic research, and thus give a new impetus to linguists and language experts to employ digital services for data analysis.

Keywords: corpus methodologies, quantitative analysis, qualitative analysis, African languages, IGT, valency

1. Introduction

For African languages the need for corpus creation is paramount. Two of the important desiderata are that these resources become widely accessible, also in Africa, and that they be structured such that they can be easily used for community-driven language development. The resources we would like to present here lend themselves to this purpose, in a methodology of channeling various types of commonly available resources into combined repositories where we not only can represent the existing resources on a common platform, but also use them for producing more advanced views of data and constellations of data of African languages. The 'mother' structure of this methodology is TypeCraft, which is specialised on Interlinear Glossed Text (IGT), that is, corpora of manually morpheme-to-morpheme annotated natural language text. Modules with which it cooperates include the LCC (Beermann et al 2016), which offers monolingual corpora of standard sizes from different sources such as the web, newspapers and the Wikipedia, and a multilingual valency project which in this case uses lexical resources from a Toolbox project (underlying Kropp Dakubu 2009). Such a combination of different resources is not unusual in the work with lesser-resourced languages, and not without problems (Beermann and Bouda 2014), but it introduces new possibilities for the local communities as already existing small resources can be combined. This is essential for linguistic research. The question we will focus on here is how IGT can be explored for linguistic purposes using annotation mining, and how valence information can be made accessible online by augmenting interlinear glossed texts.

The TypeCraft (TC) application (Beermann & Mihaylov, 2014) is an open infrastructure that allows for the creation and retrieval of IGT data - the standard data format in linguistics. TC is a user-driven database. Its main function is to enable the sharing of linguistic data, such as transcribed and annotated oral narrations, annotated small texts, and linguistic collections exposing phenomena of special interest to linguists, such as *multi-verb constructions*, *valence frames*, *tense-aspect systems*, *infinitival* and other *hypotactic construction* types (to just name some). At present TC hosts 2137 texts from 146 languages (for an overview see Table 1).

The database comes with several data management tools, such as linguistic editor, a text importer, an exporter, a collaborative editing tool and a search facility with a graphical menu-based interface. It is the latter which figures centrally for annotation mining. For our presentation we have chosen to represent data from Akan.

Data type	*Data count*
Text count	2145
Phrase count	316,604
Word count	5,297,405
Morpheme count	4,527,478
Part-of-speech tagged words	4,851,807
Gloss-tagged morphemes	330,714
Sense-tagged morphemes	1,173

Table 1: TypeCraft database in terms of stored data and annotations assigned.

Using standard query techniques and simple means of data visualisation, we will use data from Akan and Ga, both Kwa languages (ISO-693-aka, ISO-693-gaa) soken in Ghana to show how IGT corpus data can inform linguistic research. Figure 1 for example shows the absolute number of the most important Akan gloss tags, while Figure 2 shows the distribution of Akan part of speech tags. In our presentation we will, e.g. discuss linguistic patterns arising from the cross-classification of this information.

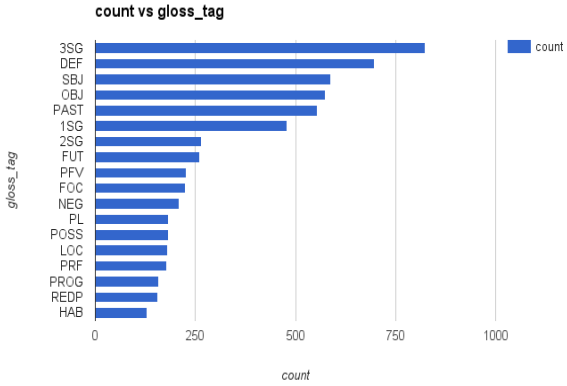

Figure 1 Absolute numbers of the most frequent gloss tags in the TypeCraft Akan corpus

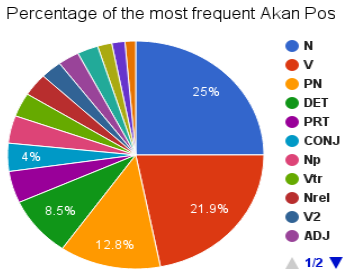

Figure 2 Percentage of the most frequent Pos tags in the Akan TypeCraft corpus.

We furthermore will be able to show that our overall set-up allows us to extend our resource efficiently, and to prepare them for linguistic use where quantity of data counts as much as a certain depth of annotation. We will exemplify this with the construction of valency resources, a type of resource almost absent for LRLs, despite a rather rich flora of valency lexicons, valency banks, etc., for well resourced languages. For an LRL, building a complete valency resource from scratch may well be out of question, but we will show that it is in principle fully possible to approach such a matter in a stepwise fashion, in tandem establishing valency frame types, finding examples of sentences instantiating them, entering them in an online database with an user-friendly online interface. Crucial is that this be feasible as a community project among linguists and language experts, which requires transparency in classification and labeling, ease in access to the classification interface, and ease in data search and retrieval of larger scale regularities. These are desiderata which TC is able to meet. TC also allows for the scaling up of information building, be it through pattern based generalizations over data already in the database, or import of data from other sources, in such a way that the format of the imported data matches that of the existing data. TC furthermore allows for a common classification of valence information across languages, whereby profiles of inventories of valency frame types can be made accessible analogously to what was shown in Figures 1 and 2 for GLOSS and POS categories, and cross-language comparisons can be made.

Our main examples for our presentation will be Ga and Akan, both languages being linguistically well researched but both still without substantial and available digital resources We will present our resources from Ga which are derived from Kropp-Dakubu's paper dictionaries and private digital resources. For Akan there are larger corpora of written and transcripbed spoken text some of them with in-depth annotation and curated. We then will describe how a cluster of valency related resources have been developed for Ga and instantiated in TC.

Steps in the development of valency resources for Ga.
1. Linguistic and lexicographical work establishing grammar descriptions of Ga and dictionaries of Ga, the latter supported by a Toolbox project for Ga.
2. Manual construction of a classificatory overview of valency frame types in Ga.
3. Manual generation of sample sentences in TC with annotation for valency frames, coded in a formalism specifically designed for valency information, in addition to standard IGT.
4. The development of a monograph, Dakubu (unpubl), so far unpublished, describing the valency frames of a large number of Ga verbs.
5. Expanding the Toolbox project adding systematic valency descriptions for its verbs, using the same coding formalism.
6. Importing (via some interesting but here irrelevant steps) the information from step 4 into the online database MultiVal,[1] where valency information for verbs from four languages is represented in a common code, and comparative search is thus possible.
7. (Projected) Importing the information from step 5 into TC directly, using the same format as in step 3, and thus supplementing the information already available in TC from step 3.
8. (Projected) Importing annotated examples directly from corpora into TC, employing generalization and

[1] Cf. Hellan et al. 2014, and https://typecraft.org/tc2wiki/Multilingual_Verb_Valence_Lexicon.

XML technology based on information accumulated through steps 3 and 7.

The code developed at stage 2 for valency annotation is the system *Construction Labelling (CL)* (Hellan and Dakubu 2010, Dakubu and Hellan 2017, and at https://typecraft.org/tc2wiki/Ga_Valence_Profile.). The CL valency annotation 'templates' are written as illustrated in (1), applicable to a sentence like (2).

(1) v-tr-suAg_obTh-CREATION
(reads: "a verb-headed transitive syntactic frame where the subject carries an agent role and the object a patient role, and the situation type expressed is CREATION")

(2) **E-fee** **flɔɔ**
3S.AOR-make stew
'she made stew'

Step 3 uses this code, in a small corpus shown on the TC wiki https://typecraft.org/tc2wiki/Ga_annotated_corpus; see illustration in Figure 3 below. A preliminary version of the monograph representing step 4, which uses the code, is available at https://typecraft.org/tc2wiki/Ga_Valence_Profile.

Step 5 utilizes the Toolbox project used as basis for the general purpose dictionary (Dakubu 2009). All lexical entries are exemplified by standard annotated examples; for verbs the exemplifying expressions are sentences. In the step 5 augmented Toolbox edition, all verb entries are systematically annotated for *valency* such that each entry reflects a unique valency frame, as exemplified in Table 2, for the verb *fee* as used in (2); the valency codes are written into the lexical entry following the general 'field' style of Toolbox, here as the fields \sl1 (POS of head), \sl2 (valency frame), \sl4 (thematic roles), \sl6 (situation type):

Table 2 Example of Toolbox entry enriched with CL valency annotation

\lx **fee**
\hm 2
\ph fèê, fèé, !fé \ps verb
\sn 1 \ge make \de make, do, perform
\sl1 v
\sl2 tr
\sl4 suAg_obTh
\sl6 CREATION
\xv E-fee flɔɔ, samala
\xg 3S.AOR-make stew
\xe she made stew, soap.

While step 6, the import into the MultiVal representation, is a separate track to the one involving TC, and concerns only valency marking, not IGT, it shows that the valency aspects of the specifications in the Toolbox version can be readily exported to other formats. Moreover the content of the information in question is already part of the TC valency representation format, both in smaller corpora and in larger ones,[2] so that the MultiVal import is a good preparation for step 7. In particular an intermediate formalization stage, described in Dakubu and Hellan (submitted), where 547 verb lexemes receive altogether 2006 entries due to many verbs having multiple frames, allows one to already formulate regularities as to valency classes, such as which frame types tend to combine for how many verbs, and more, constellations for which TC can in principle offer a search interface. Through the supplement of such a resource with a valency annotated corpus, one then has the ingredients of a full-fledged valency lexicon.
Step 7 will offer the advantage of jointly specifying IGT and valency, a combination of information necessary to fully appreciate the valency specification, and on the other hand a valuable supplement to the standard IGT.[3] To illustrate, for a specification such as the one in Table 2, TC will offer the specification in Figure 3:

String: Efee flɔɔ
Free translation: She made stew.

Morph	|E	|fee	|flɔɔ
Citation	|	|make	|stew
GLOSS	|3.SING	|AOR	|
POS	|V		|N

Efee: SAS: NP+NP
FCT: transitive
ConstructionLabel: v-tr-suAg_obTh-CREATION

Figure 3 Integrated format for IGT and valency information in TC

The standard TC importer will produce the upper part of the specification, and a valency conversion code also used in MultiVal will expand the fields sl1, sl2, sl4, sl6 into the valency display with SAS (for 'syntactic argument structure'), FCT (for 'functional label') and 'ConstructionLabel'. The number of sentences involved at this stage will be around 2000, corresponding to the example sentences offered in the Toolbox version, based on the field specifications under \xv, \xg and \xe in all its entries, exemplified in Table 2.
With step 7 done, a significant amount of data is in place for attempting automatic induction from corpora, i.e., step 8. Already for IGT (with GLOSS information being more difficult than POS information) a research question is how much already annotated data must be in place in order for automatic acquisition from 'raw' text to be

[2] See for instance https://typecraft.org/tc2wiki/Norwegian_Valency_Corpus, using text data from the LCC and with valency and IGT data induced within TC with use of digital resources for Norwegian.
[3] By the latter, the IGT is a step closer to a syntactic treebank, insofar as valency is an organizing factor of syntactic structure, however without making any commitments to syntactic framework. By the same token, IGTs so organized are possible inputs to Grammar Induction algorithms, as described in Hellan and Beermann 2014.

possible. For valency this will be even more a research topic, for which TC will provide a good ground, allowing access to IGT in the creation of hypotheses for valency identification.
The scarceness of digital text resources for Ga, on the other hand, may provide limitations to this step. For Akan, on the contrary, the TC digital text resources are fairly rich, and the question is to what extent a similar course of actions could be built up for Akan. Here, TC has in depth annotated IGT data, but much less for valency. A question here will be whether automatic procedures could 'borrow' information from Ga, which is a close relative of Akan within the Kwa family and with many attested morpho-grammatical similarities. With due provisos concerning automatic valency addition to the existing IGT, but keeping in mind the possibility of manual annotation, one might well achieve an interesting corpus of valency annotated sentences also for Akan, and if so, strategies for automatic induction of valency from digital text will be a possibility for Akan just as much as for Ga.

Most importantly our infrastructure allows us to give African linguists and language experts direct working access to data from their languages, and to extend existing resources by resources that can be developed and customised according to individual and community needs.

References

Beermann, D., Mihaylov, P. (2014). TypeCraft collaborative databasing and resource sharing for linguists. Lang. Resour. Eval. 48, 2 (June 2014), 203-225. http://dx.doi.org/10.1007/s10579-013-9257-9.

Beermann D., and Peter Bouda. (2014). Using GrAF for Advanced Convertibility of IGT data. LREC 2014, Ninth International Conference on Language Resources and Evaluation.

Beermann, Dorothee; Hellan, Lars; Quasthoff, Uwe; Eckart, Thomas; Kuras, Christoph; Haugland, Tormod. (2016). Quantitative and Qualitative Analysis in the work with African Languages. I: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association.

Dakubu, M. E. Kropp, 2009. *Ga-English Dictionary with English-Ga Index.* Accra: Black Mask Publishers

Dakubu, M. E. Kropp. Unpubl. Ga Verbs templated.

Dakubu, M.E. Kropp and Lars Hellan, 2017. A labeling system for valency: linguistic coverage and applications. In Hellan, L., Malchukov, A., and Cennamo, M (eds) *Contrastive studies in Valency*. Amsterdam & Philadelphia: John Benjamins Publishing Co.

Dakubu, M.E. Kropp and Lars Hellan (submitted) Verb Classes and Valency Classes in Ga. Based on presentation at SyWAL 2016, Vienna.

Goldhahn, D., Eckart, T., Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012.

Hellan, Lars and M.E. Kropp Dakubu, 2010. *Identifying verb constructions cross-linguistically*. In *Studies in the Languages of the Volta Basin* 6.3. Legon: Linguistics Department, University of Ghana.

Hellan, L., Beermann, D. (2014). Inducing grammars from IGT. In Z. Vetulani and J. Mariani (eds.) Human Language Technologies as a Challenge for Computer Science and Linguistics. Springer.

Hellan, L., D. Beermann, T. Bruland, M.E.K. Dakubu, and M. Marimon (2014) *MultiVal*: Towards a multilingual valence lexicon**.** In Calzolari et al. (eds) 2014.

Mohr, G., Kimpton, M., Stack, M., Ranitovic, I. (2004). "Introduction to Heritrix, an archival quality web crawler" (PDF). Proceedings of the 4th International Web Archiving Workshop (IWAW'04).

Petrov, S., Das, D., McDonald, R. (2012). A Universal Part-of-Speech Tagset. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey.

Soehn, J-P, Zinsmeister, H., Rehm, G. (2008). Requirements of a User-Friendly, General-Purpose Corpus Query Interface. Proc. of the LREC 2008 Workshop "Sustainability of Language Resources and Tools for Natural Language Processing", Andreas Witt, Georg Rehm, Thomas Schmidt, Khalid Choukri, Lou Burnard (eds.).