Convergent development of digital resources for West African Languages

Dorothee Beermann, Lars Hellan, Tormod Haugland
NTNU
N-7431 Trondheim, Norway
{dorothee.beermann@ntnu.no, lars.hellan@ntnu.no, tormod.haugland@gmail.com}

Abstract

We describe existing resources of the Kwa languages Akan and Ga, with a view to transfer of resources well developed for one to the other. While we can build on an Interlinear Glossed Text (IGT) corpus for Akan we have a modern digital lexicon for Ga, something we still lack for Akan, while we have only very limited IGT data for Ga. While it is normally the case that annotations from a resource rich language are transferred to a resource poor language, we are here preparing our resources to allow for a transfer approach between two resource-low but closely related languages. We envisage this to be a viable strategy also for other pairs of closely related under resourced languages.

Keywords: Akan, Ga, Interlinear Glossed Text, valence lexicon, morphological tagging, transfer learning between two resource-low but closely related languages.

1. Introduction

Akan and Ga are Kwa languages spoken in the southern and south-western parts of Ghana, and two of its official languages. Akan (ISO-639-3 "aka") is spoken by about 8 million native speakers according to the LDC[1]. The language has been studied extensively over many years (publications dating at least back to Christaller 1875, 1881), yet it still lacks most of the basic digital language resources, such as a lexicon, corpora, morphological analysers, and taggers. Ga (ISO-693-3 "gaa") is spoken mainly in the Accra area by about 745 000 speakers, according to Ethnologue[2]. It also has a literature dating back many years, starting with Rask (1828), and like Akan it lacks the basic digital resources, with one noteworthy exception, viz. a modern dictionary, compiled by Mary Esther Kropp Dakubu (Dakubu 2009), an authority in the study of West African languages and an expert of the language.

Having access to linguistic resources from two closely related Kwa languages, the line of research that we are interested in is driven by the question whether convergent development of closely related under-resourced languages, such as Akan and Ga, can create an opportunity to develop the basic digital resources for both languages more efficiently. In NLP, transfer learning is used as a methodology whereby resources from a resource rich language are transferred to a resource poor language. Can a similar approach be used whereby a digital resource from a poor resource language is transferred to a closely related resource poor language? In this paper we present our digital resources for Akan and Ga, which consist of an Interlinear Glossed Text (IGT) repository and a morphological tagger for Akan, and a digital valence lexicon for Ga, in the light of this question.

In section 2 we describe the curation of an Akan corpus and the development of a morphological tagger for the language. In both cases we combine community driven manual annotation with the automatic parsing of our IGT resources. In section 3 we describe, for Ga, the digitalization of a Toolbox lexicon and its conversion to a valence lexicon. We consider the learning from lexical data in the context of the semi-automatic valence annotation of Ga and eventually also of Akan. One of our long-term goals is to advance parsing for Akan using Ga resources, and the use of automatic annotation procedures for a more efficient enlargement of our West African IGT corpora.

2. Akan

Our Akan corpus consists of 102 IGT-style annotated texts, mostly linguistic sentence collections and small transcribed oral narratives. The corpus was created using a collaborative approach. G*raduate students* were asked to work on class projects which involved the morpho-syntactic *annotation* of their native language.

TypeCraft Akan resources	Words	Phrases
TypeCraft owned Akan resources	28 429	2689
TypeCraft hosted Akan resources	96 697	7535

Table 1: Snapshot of the TypeCraft Akan corpora

1 Linguistic Data Consortium, https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/west-african-languages.pdf (accessed 21.01.2018)

2 Ethnologue, https://www.ethnologue.com/language/gaa (accessed 21.01.2018)

For our work we used the *TypeCraft* research tool,[3] which contains two different sub-corpora for Akan (see Table 1), one sub-corpus consisting of 7535 phrases annotated by native speaker students of linguistics, and one TypeCraft-owned corpus consisting of 2689 phrases, which builds on the Akan data that was hosted at TypeCraft. In the case of the former, we were granted the necessary permissions by the owners (thus, for instance, for graduate work at our University, the students' consent was sought for use of their work in further research). In order to systematize our work with the TypeCraft owned corpus, we in 2016 started an Akan corpus curation project which in its first phase undertook the manual re-annotation of the TypeCraft-owned Akan data. At the same time, we started to enlarge that corpus more systematically. The Akan data hosted on TypeCraft was not affected by the effort, as that data is owned by individual TypeCraft users. The curation effort was accompanied by phonetic studies of Akan Tone.[4] In the project's first phase we re-annotated 2689 phrases manually. We followed a community approach receiving the help of

2.1 Annotation Profiling

Figure 1 shows a comparison for the most frequently assigned gloss tags for Akan. To the left we show their ranking for 2015 and one right for 2018. The 2015 snapshot was taken for all Akan data then hosted by TypeCraft. The snapshot from 2018 was performed on our own Akan data. The gloss profile from 2015 still reflects the work of student annotators who were native speakers of the language. The students did not receive special annotation training as part of their linguistic studies, so that they were informed but not supervised annotators. The 2018 annotation profile reflects the work of expert linguists working together with trained student annotators. Figure 1 shows several things:

(I) The 6 most frequent tags remain the same for the 2015 and the 2018 corpus, although for all of the labels their absolute numbers and their ranking relative to each other may have changed, as noted under (II).
(II) The categorization of features for the verbal inflection has been reconsidered; an exception is the assignment of past tense which is in both 2015 and 2018 the most

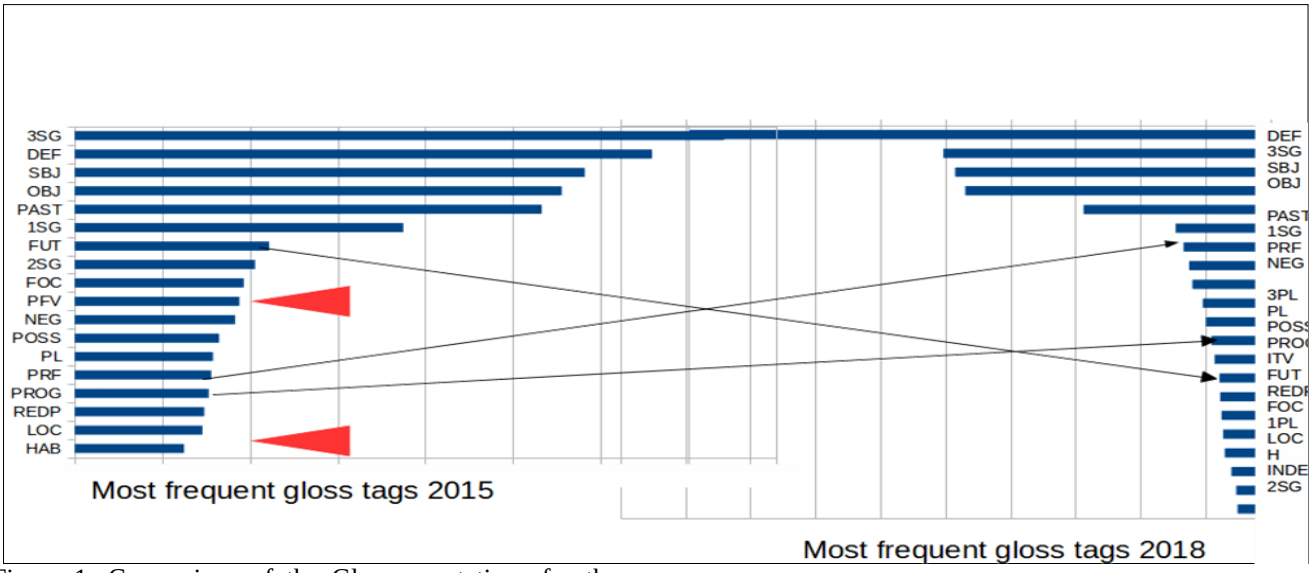


Figure 1: Comparison of the Gloss annotations for the 2015 and 2018 version of the TypeCraft Akan corpus

the Ghanaian Student Association at our university as well as the support of Akan speaking senior linguists. In addition, we hired Akan students as part-time data analysts.
To monitor conciseness and consistency of our annotations, we continuously used Annotation Profiling, a methodology based on the analysis of words and morphemes bound annotations. We will describe this effort in the next section.

frequently assigned tense label. As indicated by the small black arrows in Figure 1 the ranking of the tense and aspect features future, perfect and progressive has changed. While the morphological marking is unambiguous, that is, the Akan prefix bɛ- stands for future, *a-* for perfect (unless the verb is negative), and *re-* for the progressive, prior annotations tend to reflect the tense expressed in the English translation of the sentence rather than the actual value of the Akan morpheme.
(III) Concerning again verbal features, the perfective and the habitual which figured prominently in the 2015 annotations (see red arrows), are no longer between the most frequently tagged grammatical features, as we rectified errors which for the most frequently assigned tags listed in Figure 1 concerned the difference between the perfective aspect and the perfect tense. The difference is not easily pinned down, especially when sentences appear in isolation, as perfect verb forms can have a perfective meaning. Example (1) illustrates what is meant. The sentence describes a scene

3 TypeCraft (https://typecraft.org) is a service. It can be used online by individual users and projects. As a service TypeCraft *hosts* data. The TypeCraft project is a research group which as one of its activities curates data using the TypeCraft application. The data provided by the TypeCraft project is Typecraft *owned* data.

4 This has developed into a sub project in its own right, cf. Van Dommelen and Beermann (forthcoming).

in a video clip where a cat is looking at a man for a while without him waking up. And in fact the *a*-prefix on the verb *hwɛ* meaning 'look' expresses the perfect tense (PRF), not the aspectual perfective (i.e., completed aspect, marked PFV, as wrongly marked in (1)). (Whether Akan is predominately an aspect or a tense marking language is a long standing discussion in Akan studies (Dolphyne (1988, 1996), Boadi (2008), Osam (1994)).)

(1) Wahwɛ ara nso still papa no nsɔre.

w	a	hwɛ	ara	nso	still	papa	no	n	sɔre
3SG	PFV	*look*	FOC	FOC		*man*.SBJ	DEF	NEG	*get_up*
V			PRT	PRT		N	DET	V	

"It looked (for some time) but still the man is not getting up"

Generated in TypeCraft.

One can say that the expert annotation led to an increased depth of annotation for all parts of the grammar, especially however for the verbal inflection:
(A) Preverbs such as spatial verbs serving as inchoative markers, which were mostly not annotated in 2015, now received an annotation, in Figure 1 reflected by the tag: ITV 'itive'.
(B) multi-functional formatives where now annotated in context, which in Figure 1 is reflected in a decrease in the formatives classified as focus markers, which to us seems to appear as a label when one was not so very sure what the grammatical function really was.
(C) In 2015 mainly, definite nominal modifiers were identified, now also indefinite modification is tagged.
(D) The coverage for negation and relational nouns was increased. In Figure 1, the tag LOC mainly points to relational nouns which are tagged as: POS: Nrel, Gloss: LOC.

In summary our curation effort resulted in the improved conciseness of our annotations especially for the coverage of the verbal inflection; much more work needs to be done for the nominal system. We also improved the consistency of annotations and achieved more depth in annotation.
Finally, for the evaluation of our results we also used trailing annotation profiles as heuristics (see Figure 2). To start with, trailing annotations pointed to random tags which where only assigned once or twice, such as ACC (accusative), or ADD (additive aspect) for cases of reduplication. In our present corpus, we still find trailing annotation contours with over 50 tags, however, these reflect that some

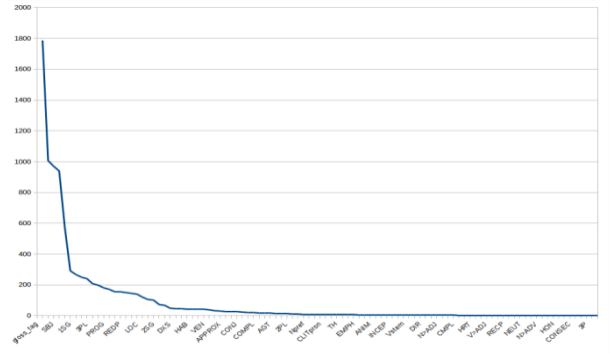


Figure 2: Trailing annotation contour

annotators chose to annotate aspects of the grammar which were not yet targeted for annotation by the project, such as reference, derivational morphemes and thematic roles.

2.2 Extending the corpus and automatic tagging

During the second phase of the curation project, we added with the radio corpus a new resource to our Akan corpus. It consists of 10 texts of between 60 and 100 sentences of transcribed and translated radio conversations between a Ghanaian radio host and his guests.[5] The material reflects contemporary spoken Akan, and prominently features code switching between Akan and English by all speakers.
In parallel, we had worked for a while on the development of an Akan tagger. In the first cycle of our curation project we had trained and tested the tagger on the material we had curated, and we proceeded to test it on one the radio texts, which was then unseen data. The challenges for the parsing of this newly acquired corpus resided in two factors. Although the radio corpus was large compared to our other IGT resources, under testing we still had to deal with the scarceness of data. Secondly, while our training data had no codeswitching, the radio corpus was a codeswitching corpus, that is native Akan speakers were alternating between Akan and English.
The tagger uses a hybrid approach to tagging for both Part-of-speech and gloss tags. It is primarily a universal context-processor, which translates a parsed and annotated sentence on the word and morpheme level into a set of context features. The feature set used by the tagger is configurable. While most taggers use a rather sparse feature set (cf. Schmid 1995, Toutanova et al. 2000), we use a rich feature set of up to several hundred distinct types of features. When supplied with training data, the tagger extracts, according to configuration, all context features observed, and stores them in a database to create a language model. When tagging on untagged data, context features are in the same manner extracted in a left-to-right fashion. That is, at word *n* we have information about the surrounding words (and possibly morphemes) and the *n-1* preceding inferred tags. These contexts are then matched with a set of tags (per context) which are assigned a probability based on their likelihood. The probability generation is performed using Bayesian inference based on occurrence count of the context. Some adjustments are made however. We adjust the feature probabilities upwards for features that are complex, and downwards for features that are rarely seen. The most probable Bayesian estimator is then selected.

On completion, the tagger iteratively reruns the tagging procedure a configurable number of times. On these subsequent runs the features available should now be richer, the idea being that the tagger should iteratively correct its own mistakes when supplying itself with more context.

5 The radio shows were recorded and transcribed by S.Brobbey 2015.

Secondarily the tagger can be configured with specific definite rules which map (a combination of) context features directly to tags. This allows the tagger to deal with noisy training data by correcting the generated language model with overriding rules. For the Akan language model, several such rules were incorporated.

We parsed the corpus using an English and an Akan language model, a process that we will not describe further here. The results were poor for our first run of one of the radio corpus texts, as shown in Table 2.

	Precision	Recall
POS tags	0.72	0.72
Gloss tags	0.70	0.80

Table 2: Classification measures for unseen Akan data.

To improve the performance, we manually re-annotated that text and re-trained the tagger again using these 60 sentences long text.
We further noticed that our effort put in the manual re-annotation did give us some improvement in precision and recall, most likely due to the reduction of inconsistency, but still left us with a flawed linguistic representation of Akan. So taking everything into consideration, in spite of a further round of re-annotation we still had noisier training data than normally is used for the creation of annotated corpora. The use of noisy trainings data is also described by Garrette and Baldridge (2013), who focused on POS tagging using 14 different tags. We dealt in our project with a considerably larger number of word and morpheme level tags, which then also meant a higher and several sources for the inconsistency of the annotation in our training data. In order to arrive also at a grammatically adequate corpus of Akan, we needed to implement on top of Bayesian inference a set of conditions reflecting the basic rules of the Akan grammar. For the present tagger development, we focused on the verbal inflection, and some very basic syntactic rules concerning the position of nouns and their modifiers. With all this in place we re-ran the parser. The considerably improved results are shown in Table 3 and 4, once with direct mapping as our rules enforced, and once without, again for POS and Gloss annotations.The results are calculated by weighted averages over total positives for each tag.

	Tag	**Precision**	**Recall**
Without rules	ADJ	0.93	1
	ADV	1	1
	CONJ	0.95	0.94
	DET	0.77	0.96
	N	0.95	0.99
	PN	0.91	0.96
	PREP	0.95	1
	PUN	1	1
	V	0.56	0.9
	TOTAL	0.83	0.93
With rules	ADJ	0.93	0.93
	ADV	0.78	0.91
	CONJ	0.9	0.94
	DET	0.55	0.97
	N	0.92	0.93
	PN	0.59	0.59
	PREP	0.91	0.95
	PUN	1	1
	V	0.87	0.9
	TOTAL	0.78	0.84

Table 3: Classification results for a selection of POS-tags for seen Akan data.

The "without rules" results are the tagger tagging with no assistance by overriding rules. The total score is calculated by weighted averages over total positives.
In both cases the total precision/recall/f1 ratings are calculated by weighted averages over total positives. Note that directly comparing the result-sets in Table 2 with the results shown in Table 3 and 4 may be misleading, as the improved result is on seen data, while the results shown in Table 2 are on unseen data.

	Tag	**Precision**	**Recall**
Without rules	<empty gloss>	0.98	0.89
	1PL	0.75	0.98
	1SG	0.62	0.94
	2PL	0.67	1
	2SG	0.51	0.96
	3PL	0.78	0.99
	3SG	0.75	0.88
	FUT	0.95	1
	NEG	0.89	0.95
	PROG	0.94	0.97
	TOTAL	0.87	0.86
With rules	<empty gloss>	0.98	0.87
	1PL	0.56	1
	1SG	0.6	1
	2PL	0.67	1
	2SG	0.51	0.92
	3PL	0.78	0.99
	3SG	0.75	0.9
	FUT	0.95	1
	NEG	0.91	0.96
	PROG	0.94	0.97
	TOTAL	0.86	0.85

Table 4: Classification results for a selection of Gloss-tags for seen Akan data.

The tagger in its present stage does not have built-in strategies reflecting syntactic structure of the strings processed, and no strategies reflecting valency information about the lexical items occurring, strategies which of course could add to parsing adequacy. To our knowledge there exist no IGT parsers of Akan, and no digital lexical resources which could be built into the current tagger.[6] In order to make such strategies in principle available to the development of the present tagger, we therefore will explore strategies of transferring information from our Ga resources.

6 Dictionaries like Christaller (1881) and Anyidoho (2006) are not amenable to digital employment.

2.3 Tagger configuration and evaluation

The most important configuration entries of the tagger can be found in Table 5.

Number of iterations per tagging	3
Max *n*-gram length	4
Max length of context feature combinations	3
Ignore empty POS	True

Table 5: Important configurations for the tagger. When combining context features into more complex context features,

When training for English, the configuration was slightly changed by letting the *n*-gram length and combination length be 2 and 1, respectively. The tagger was also configured with specific context feature type weighting. The base context features used (which are combined to more complex features) can be found in Table 6.

Word
Morpheme
Surrounding ngram (of words, POS, etc.)
Prefix ngram (of words, POS, etc.)
Suffix ngram (of words, POS, etc.)
Gloss
Citation form

Table 6: The context features used in training and evaluation. These feature fom the base, or atomic, feature types used, and are combined to more complex context features.

The tagger was first trained on and evaluated with Akan data. The training data was split up into 80%/20% training and test data (in total about 5000 word tokens), for which the tagger had an F1 score of 57%. It was then trained on English, primarily on direct word to tag features. It was not evaluated on English alone.

3. Ga

The starting point for our work with Ga is a Toolbox project holding data of the general-purpose published dictionary (Dakubu 2009). The lexicon file consists of 80,000 lines of code, with 7080 entries, of which 5014 for nouns, and 935 for verbs, of which 722 were annotated for valence. From this Toolbox repository we created a valence lexicon.

3.1 Toolbox lexicon augmented by valence information

In the Toolbox edition used, verb entries are systematically annotated for *valency* such that each entry reflects a unique valence frame. The code used in this annotation is the system *Construction Labelling (CL)* (Hellan and Dakubu 2009, 2010, Dakubu and Hellan 2017). Following the overall left-to-right order indicated in the schema in (2), the CL valency annotation 'templates' are written as illustrated in (3), with the information between each pair of slashes or underscores counting as a 'minimal construction unit' (MCU):

(2) head – valenceFrame – special properties of syntactic constituents – semantic roles of constituents – aspect, Aktionsart – situation type

(3) v-tr-suAg_obTh-CREATION

A paraphrase of (3) is: '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'.)
This template is applicable to a sentence like (4).

(4) E-fee flɔɔ
3S.AOR-make stew

'she made stew'

The design of a lexical entry in the amended Toolbox version is exemplified in Figure 2, for the verb *fee* as used in (4); the valence codes are written into the lexical entry following the general 'field' style of Toolbox, here as the fields \sl1, \sl2, \sl4, \sl6:

\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

Figure 2: Example of Ga Toolbox entry enriched with CL valence annotation

A verb with more than one valence frame has one entry specified per frame, hence the verb *ba*, for instance, is represented by 15 different entries in this edition of the Toolbox file. 547 verb lexemes here received altogether 2006 entries annotated in this fashion. In Figure 2, the specification '\hm 2' indicates that this is the second lexeme entered with the form *fee*.[7]
The above resource is also available as a lexical data structure of the type used in Head-Driven Phrase Structure Grammar (HPSG) [8] implemented grammar. The present version consists of 1980 sequentially numbered entries,

7 An overview of full CL templates established for Ga can be seen at: https://typecraft.org/tc2wiki/Ga_Valence_Profile.
8 Cf. Pollard and Sag 1994, Sag et al. 2003. HPSG uses the formalism of Typed Feature Structures (Copestake 2002), whereby every object in the grammar and lexicon belongs to a type; types are organized in multi-inheritance hierarchies.

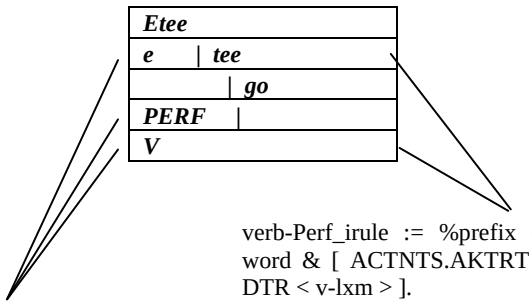
now using the style of notation in (3) in the top line of the entry to indicate the *lexical type* to which the entry belongs. The example in Figure 3 shows a direct counterpart to the Toolbox entry in Figure 2, with *fee_244* as the entry identifier (the formula part ':= v-tr-suAg_obTh-CREATION' means 'belongs to the type v-tr-suAg_obTh-CREATION'):

fee 244 := v-tr-suAg_obTh-CREATION &
[STEM <"fee">,
PHON <"fee">,
ENGL-GLOSS <"make">,
EXAMPLE "E-fee flɔɔ, samala",
GLOSS "3S.AOR-make stew",
FREE-TRANSL "she made stew, soap."].

Figure 3: HPSG style counterpart to the entry in Figure 2

3.2 Inferring IGT from HPSG type lexical data

It is possible to exchange information between IGT and HPSG grammars. A way of inferring information for an HPSG grammar from IGT is illustrated in Figure 4, this approach is described in Hellan and Beermann (2014) with exemplification for Ga; the implementation framework itself is called *TypeGram.*[9] Here, from a snippet of a Ga IGT like the one indicated, one can infer the grammar specification indicated underneath the snippet, being fragments of a lexical specification and an inflectional rule formulation (attributes such as 'ORTH', 'AKTRT' etc., and value categories such as *v-lxm, perf* and *word*, are defined in the general grammar system):



verb-Perf_irule := %prefix (* e) word & [ACTNTS.AKTRT perf, DTR < v-lxm >].

tee-v := v-lxm & [ORTH <"tee">, …].

Figure 4 Illustration of correspondences between IGT and HPSG grammar encoding

Inference of IGT from an HPSG grammar, including valence information, is in turn described in Hellan et al. 2017, the IGT being generated as part of the parse result. While this involves a full grammar, partial inference can also be done from parts of a grammar, such as valence information into an IGT from an HPSG type of lexicon, given recognition of lemma forms in the strings to be annotated. This will be feasible if the procedure can be combined with a morphological parser like the Context parser for Akan described in section 2. By extending such a parser to Ga, and supplementing its assignments with lexical information from the lexicon file, we hope in a future step to make the valence information from Ga operational for Akan.

3.3 Ga valence features

The lexicon file is by itself a large text file,[10] where lexical specifications and valence information are laid out as illustrated above. Of particular interest in a Kwa perspective are construction types quite common in the language but hardly found in European languages. Some types are mentioned below, with indication of the number of verb entries in which they appear as valence information, exemplified for class a and b in (5) :

a. *Bodypart relations* (158 entries)
b. *Identity relations* (110 entries)
c. Subject headed by relational noun (99 entries)
d. Object headed by relational noun (690 entries)
e. Object's specifier headed by relational noun (29 entries)

(5)
a.
v-tr-suIDobSpec_obBPobSpec-suAg_obLoc-COMMUNICATION
Ee-la e-daa-ŋ
3S.PROG-sing 3S.POSS-mouth-LOC"
V N
"He's murmuring incoherently to himself."
('suIDobSpec' = subject (expressed by a clitic) is coreferential with the specifier (expressed by a clitic) of the object;
'obBPobSpec' = object is bodypart of the specifier of the object)

b.
v-tr-obPossp_obBPobSpec-suAg_obLoc-CONTACTFORCEFUL
E-ŋmra e-toi-ŋ
3S.AOR-scrape 3S.POSS-ear-LOC
V N
"She slapped him."

The MCU spelled with capital letters is in each case the situation type to which the content of the sentence belongs; for language comparison of valence frames, such information is of course essential. The eight largest classes in the lexicon file are listed in Table 7:

COGNITION	(83 entries)
COMMUNICATION	(178 entries)
CONTACT	(56 entries)
EXPERIENCING	(45 entries)
MOTION	(180 entries)
MOTIONDIRECTED	(55 entries)
PLACEMENT	(53 entries)
PROPERTY	(164 entries)

Table 7: The most frequently used situation type labels in the Ga lexicon

At present we only have a very small annotated IGT corpus of Ga in TypeCraft, 90 phrases, however with

9 See http://typecraft.org/tc2wiki/TypeGram. For a related approach to grammar induction from IGT using LFG, see Beermann 2014.

10 Much of its information is also exposed at the online 4-languages valency lexicon *MultiVal*, cf. Hellan et al. 2014.

inclusion also of valency information along the lines here described.

3.4 Evaluation of the valence resource

The investigation of valence types in Ga can be related to the research into valency classes started with Levin (1993), followed up, i.a., in VerbNet and in the Leipzig Valency Classes (LVC) Project,[11] being attempts to associate commonalities in morpho-syntactic patterns with semantic factors, both language internally (like Levin op. cit. and VerbNet) and cross-linguistically (LVC). Establishing valency classes for Ga has a tie to VerbNet in aiming at a fairly large coverage of the language's verbs, and to LVC in establishing one more coordinate point in the attempt to attain a typologically broad basis for generalizations within this domain.

Preliminary comparisons of valency frame types for Ga and English suggest that they have less than 20% of their valency frames in common (see, e.g., Dakubu and Hellan (2017)). Even if situation types are common across languages, it is thus by no means a given that there is much commonality between languages as concerns valency classes.

Given the large discrepancies in valency frames between Ga and English, a good strategy may be to first explore commonalities between Ga and other West African languages. Some perspectives are here offered in Schaefer and Egbokhare. 2015, Creissels 2015, conducted in the frame of LVC. However, in the present setting, the natural step will be to build a mapping between Ga and Akan lexical information, assuming that the valency labels used for Ga are adequate also for Akan.

4. Conclusion[12]

With the Akan Context Tagger, we present the first IGT tagger for Akan. It has been used with homogeneous as well as with code-switching data. Our results are encouraging but further training with both types of data are necessary. We plan to use lexical information including valency information developed for Ga to increase its efficiency, which would allow us to tag larger amounts of text than what we have so far. Since the grammatical systems of the languages are not very different, and they are also not too distant lexically, integrating such information will be in principle a feasible task.

From the perspective of Ga, the extension of the parser technology for Akan to Ga should likewise be possible. An interesting issue is here whether an already small HPSG parser for Ga can be utilized in this process. This then would also allow us the syntactic parsing of both languages. From the viewpoint of research into valency classes per se, an alignment of the Ga resources with resources of Akan is desirable, but this is probably more a long-term research project than a matter of transfer of available resources, since this requires analysis at a level of detail far beyond what is required for establishing a large but basic vocabulary for efficient basic morpho-syntactic parsing.

11 Cf. for LVC, Malchukov and Comrie (eds) 2015 and http://valpal.info/; for VerbNet http://verbs.colorado.edu/~mpalmer/projects/verbnet.html.

12 We are grateful for the comments from the three reviewers of this paper.

5. Bibliographical References

Anyidoho, A. et al. (2006) Akan Dictionary. Pilot project. University of Ghana.

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

Beermann, D. and Mihaylov, P. (2014). Collaborative databasing and Resource sharing for Linguists. In: *Languages Resources and Evaluation*. Springer.

Boadi, L. (2008). Tense, Aspect and Mood in Akan. In Ameka, Felix, ed. *Aspect and Modality in Kwa Languages.* Amsterdam: John Benjamins Pub. Co., 9 – 68.

Brobbey, S. (2015) *Codeswitching on Ghanaian Radio Talk-show: "Bilingualism as an Asset".* Master's thesis, University of Bergen, Norway.

Christaller, J.G. (1875). *A Grammar of the Asante and Fante Language Called Tshi.* Gregg Press.

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

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications.

Creissels, D. (2015). Valency properties of Mandinka verbs. In: Makchukov, A., and B. Comrie (eds) Pages 221-260

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

Dakubu, M. E. Kropp. (2013). Ga Verbs and their constructions. Monograph ms, Univ. of Ghana.

Dakubu, M.E.Kropp, and L. Hellan (2017) A labeling system for valency: linguistic coverage and applications. In Hellan, L., A. Malchukov and M. Cennamo (eds) (2017).

Dolphyne, F. A. (1988). The Akan (Twi-Fante) Language: Its Sound Systems and Tonal Structure. Accra: Ghana Universities Press.

Dolphyne, F.A. (1996). A Comprehensive course in Twi (Asante) for Non – Twi Learners. Ghana Universities Press, Ghana

Garrette, D. and J. Baldridge (2013). Real-World Semi-Supervised Learning of POS-Taggers for Low-Resource Languages. *Proceedings of NAACL-HLT* 2013, pp. 138–147, Atlanta, Georgia,

Hellan, L. and M.E.Kropp Dakubu. (2009): A methodology for enhancing argument structure specification. In *Proceedings from the 4th Language Technology Conference (LTC 2009)*, Poznan.

Hellan, L. and M. E. Kropp Dakubu. (2010): *Identifying Verb Constructions Cross-Linguistically. Studies in the Languages of the Volta Basin* 6.3. Legon: Linguistics Dept., University of Ghana. http://www.typecraft.org/w/images/d/db/1_Introlabels_SLAVOB-final.pdf.

Hellan, L. and D. Beermann (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, M. Marimon. (2014). *MultiVal* – towards a multilingual valence lexicon. LREC 2014.
Hellan, L., A. Malchukov and M. Cennamo (eds) (2017) *Contrastive studies in verbal valency*. Amsterdam: J. Benjamins.
Hellan, L., D. Beermann, T. Bruland, T. Haugland, E. Aamot. (2017). Creating a Norwegian valence corpus from a deep grammar. In Vetulani (ed) *Proceedings from LTC 2017*. Poznan.
Levin, B. (1993). *English Verb Classes and Alternations.* Chicago IL: University of Chicago Press.
Malchukov, A. L. & Comrie, B. (eds.) (2015). *Valency classes in the world's languages.* Berlin: De Gruyter Mouton.
Osam, E. K. (1994). Aspects of Akan Grammar. A Functional Perspective. Ph.D. thesis, University of Oregon.
Pollard, C. and Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago University Press.
Rask, R. (1828). *Vejledning til Akra-Sproget på Kysten Ginea* (Introduction to the Accra language on the Guinea Coast).
Sag, I., Wasow, T. and Bender, E. (2003). *Syntactic Theory*. CSLI Publications, Stanford.
Schaefer, R.B, and F. O. Egbokhare. (201)5. Emai valency classes and their alternations. In Malchukov, A. and B. Comrie (eds) 2015. Pp. 261-298.
Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.
Tesnière, L. (1959). *Éleménts de syntaxe structurale*. Paris: Klincksieck.
Toutanova, K. and C. D. Manning. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70.
Van Dommelen, W. and D. Beermann (forthcoming) A study of Akan Tone – the case of NA.

6. Language Resource References

The valence resources for Ga:
https://typecraft.org/tc2wiki/Ga_Valence_Profile

TypeCraft Akan corpus:
https://typecraft.org/tc2wiki/Special:TypeCraft/PortalOfLanguages

The TypeCraft Context Tagger :
https://github.com/Typecraft/casetagger