Development of a Bambara Treebank

Ekaterina Aplonova

Higher School of Economics, Moscow
E-mail: aplooon@gmail.com

Abstract

The paper describes the development of a Bambara treebank. Bambara is a Manding language spoken in Mali. It has a corpus with morphological annotation, the creation of a treebank is the next step in Bambara corpus building. The annotation scheme is based on the Universal Dependencies model, a project providing a cross-linguistic syntactic annotation for different languages. In the main part of the paper, data conversion issues are discussed. The main challenge here is the choice of part-of-speech tags from a closed universal list. The main result of our work is an annotation guide and treebank itself, which are both available online. The Bambara treebank is integrated into the Bambara Reference Corpus and available at its site.

Keywords: treebank, syntactic annotation, Bambara, corpus building

1 Introduction

An idea to create a Bambara treebank emerged about two years ago during the discussion about development of Bambara Reference Corpus.[1] Our working group included Valentin Vydrin, Kirill Maslinsky, Anna Ivanova and Egor Solonovich. We chose Universal Dependencies as an annotation scheme, as we were looking for a model of syntactic annotation for the Bambara corpus, and the Universal Dependencies was the best-known project in the respective field. At that time, our team had no appropriate annotation tool. In September 2017, I met Francis M. Tayers, who was developing an annotation tool exactly for Universal Dependencies, and we started working together. By the end of the 2017, we got a first version of the annotation scheme for Bambara and a small treebank (approx. 13k tokens), which was presented on TLT16[2] in Prague (Aplonova, Tayers 2018). In this article, I will discuss the conversion process in more detail with a particular attention to the parts-of-speech tags. I will introduce newly created annotation guidelines and describe the integration of the treebank into the architecture of the Bambara Reference Corpus.

The paper is organized as follows. In Section 2, general information about Bambara is presented. Section 3 describes the conversion process. Section 4 shortly describes the annotation tool and annotation guidelines. Finally, Section 5 gives an overview of the syntactically annotated subcorpus of the Bambara Reference Corpus.

[1] http://cormand.huma-num.fr/
[2] https://ufal.mff.cuni.cz/tlt16/

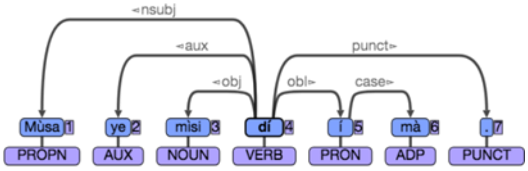
2 Bambara

2.1 General information

Bambara is the most widely-spoken language of the Manding language group (Western Mande < Manding < Niger-Congo). It is spoken mainly in Mali by 13–14 million people. Besides French, it is the major language on Malian radio and television, there are periodicals in Bambara, it is broadly used in literacy programs and in primary schools.

Bambara is a tonal language with two tones.[3] It has a rather scarce morphology. The word order is fixed: S AUX O V X, where S stands for subject, AUX for an auxiliary, O for a direct object, V for a predicate and X for an oblique. Therefore, in (1),[4] *Mùsa* is a subject, *ye* is an auxiliary, *mìsi* is a direct object, *dí* is a verb, *í* is an oblique followed by a postposition *mà*.

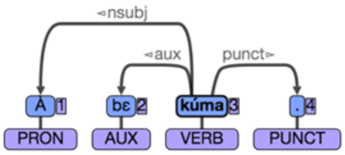
(1)



'Musa gave you a cow'.

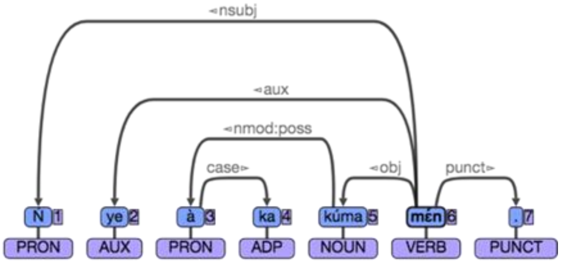
Conversion is highly productive, for example verb > noun (2), noun > adjective (3).

(2a)



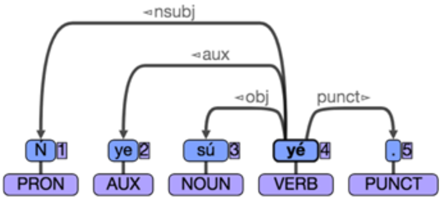
'He speaks'.

(2b)



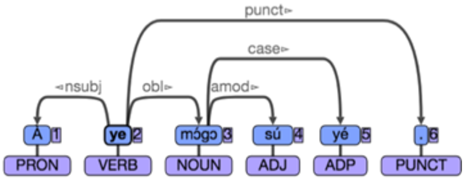
'I heard his speech'.

(3a)



'I saw a corpse'.

(3b)



'He is a useless man'.

[3] Generally, tones are not indicated in texts in Bambara. The only exception are texts transcribed by linguists.
[4] Examples are exported from UD Annotatrix tool (cf. 4). Arrows illustrate syntactic relations, beyond each token there is its part-of-speech tag. Tokens in bold are roots, i.e. heads of a whole sentence.

2.2 Bambara Reference Corpus

Bambara Reference Corpus (BRC) is available online since April 2012 (Vydrin 2013). It is composed of texts of different genres: periodicals, manuals, religious publications, texts recorded and transcribed by researches etc. Since the Bambara orthographic standard is relatively undeveloped, the corpus assumes different levels of orthographic normalization. Texts have morphological annotation[5] based on morphological analyzer Daba (Maslinsky 2014).

The corpus contains three subcorpora:

- Corbama-net: manually annotated subcorpus, where it is possible to make queries with or without tonal notation
- Corbamafara/Corfarabama: parallel subcorpora (Bambara-French or French-Bambara)
- Corbama-ud: syntactically annotated subcorpus (treebank)

3 Data conversion

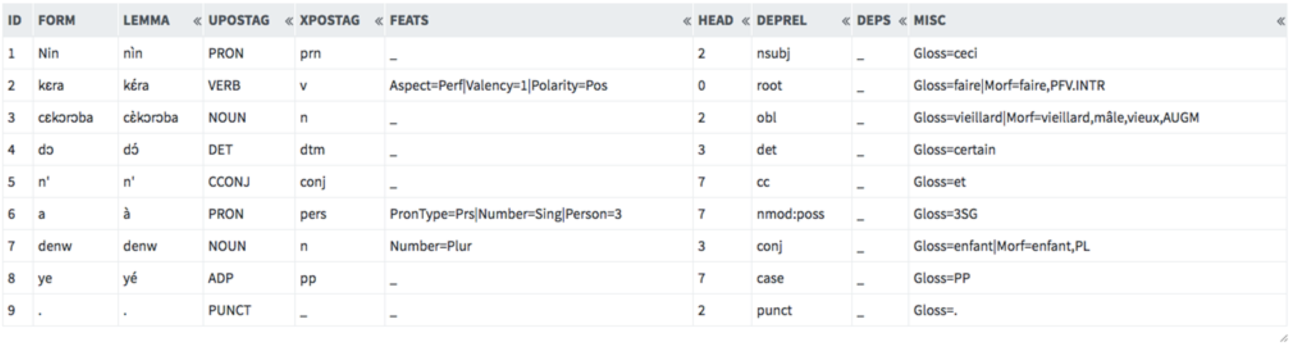
3.1 CoNLL-U

Universal Dependencies (UD) is a project that is developing cross-linguistic treebank annotation. In order to annotate syntax using UD annotation scheme, one needs to convert text, already annotated beyond the part-of-speech level, to a CoNLL-U format. In this format, sentences consist of one or more word lines, and word lines contain the following fields:

- ID: word index
- FORM: word form or punctuation; in case of Bambara, in this field, there is a token in the original orthography.
- LEMMA: lemma or stem of word form; in case of Bambara, in this field, there is a token in the standardized orthography.
- UPOSTAG: part-of-speech tag in the UD format (cf. 3.2).
- XPOSTAG: part-of-speech tag in the BRC format (cf. 3.2).
- FEATS: morphological features from the UD list[6] and/or language specific features.
- HEAD: head of the current word, which is either a value of its ID or zero for the root of a sentence.
- DEPREL: syntactic relation to the head (cf. 4).
- DEPS: enhanced dependencies.
- MISC: any other annotation

[5] The majority of texts are disambiguated automatically (approx. 4 million). Manually disambiguated texts (approx. 1 million) compose a subcorpus corbama-net, which is a «golden standard» for the parser.

[6] http://universaldependencies.org/u/feat/index.html



ID	FORM	LEMMA	UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS	MISC
1	Nin	nìn	PRON	prn	_	2	nsubj	_	Gloss=ceci
2	kɛra	kɛ́ra	VERB	v	Aspect=Perf|Valency=1|Polarity=Pos	0	root	_	Gloss=faire|Morf=faire,PFV.INTR
3	cɛkɔrɔba	cɛ̀kɔrɔba	NOUN	n	_	2	obl	_	Gloss=vieillard|Morf=vieillard,mâle,vieux,AUGM
4	dɔ	dɔ́	DET	dtm	_	3	det	_	Gloss=certain
5	n'	n'	CCONJ	conj	_	7	cc	_	Gloss=et
6	a	à	PRON	pers	PronType=Prs|Number=Sing|Person=3	7	nmod:poss	_	Gloss=3SG
7	denw	denw	NOUN	n	Number=Plur	3	conj	_	Gloss=enfant|Morf=enfant,PL
8	ye	yé	ADP	pp	_	7	case	_	Gloss=PP
9	.	.	PUNCT	_	_	2	punct	_	Gloss=.

Figure 1: A sentence in CoNLL-U.

To convert BRC to CoNLL-U, we used a Python script[7] which reads the HTML files from the BRC and performs the substitution of tags.

3.2 Part-of-speech conversion

Table 1 presents correspondences between BRC part-of-speech tag to those from UD annotation scheme.

Description	BRC tag	UD tag
Adjective	adj	ADJ
Adverb	adv	ADV
Numeral	num	NUM
Noun	n	NOUN
Proper noun	n.prop	PROPN
Verb	v	VERB
Qualitative verb	vq	VERB
Participle	ptcp	VERB
Copula	cop	VERB
Personal pronoun	pers	PRON
Pronoun	prn	PRON
Predicative marker	pm	AUX
Conjunction	conj	CCONJ
		SCONJ
Postposition	pp	ADP
Determiner	dtm	DET
Particle	prt	PART

Table 1: Conversion table for part-of-speech tags.

3.2.1 VERB challenge

In the majority of cases in Table 1, there is a direct correspondence between tags (adjective, adverb, numeral etc.). However, UD tag VERB corresponds to verb, qualitative verb, participle and copula in BRC.

In Bambara, verbs are divided into two classes: dynamic verbs (4a) and qualitative verbs (4b). Unlike dynamic verbs, qualitative verbs' predicative markers cannot express TAM values.

[7] The script is available here: https://github.com/KatyaAplonova/UD_Bambara

(4a) *À bɛ/ye/bɛ́nà kúma.* 'He speaks/spoke/will speak'.

(4b) À ka/*bɛ/*ye/*bɛ´ na` ɲìn. 'It is fine'.

They expresses attributive meanings, e.g. *ɲìn* 'be good', *kán* 'be equal', *bìlen* 'be red' etc.

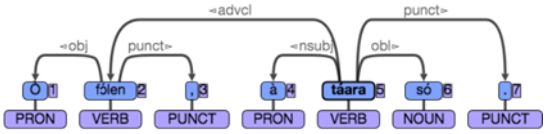
Bambara has three participles and one converb. Their functions are illustrated in the Table 2 from (Vydrin 2017).

Syntactic position	Converb	Resultative participle	Privative participle	Potential participle
predicative	+	+	+	-
attributive	+	+	+	+
nominal	-	+	+	+

Table 2: Participles in Bambara.

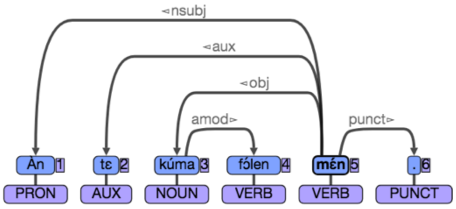
As one can see, the converb and the participles can appear in very different syntactic contexts.

(5a)



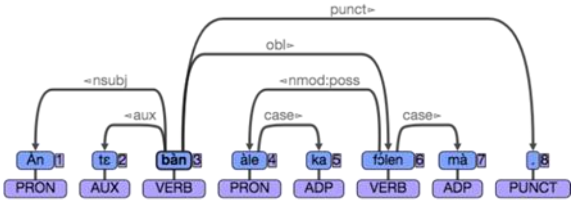
'Having said this, he left home'.

(5b)



'We do not hear the pronounced speech'.

(5c)



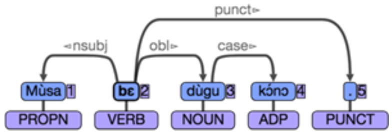
'We do not contest what he said' (his words).

In (5), resultative participle is used as a predicate (a), as an attribute (b) and as a subject (c).

The fact that participles are inevitably annotated as VERBs might lead to a misunderstanding of annotation scheme. In Section 2.1, conversion was discussed. As one can see from (2, 3), in case of conversion, we chose syntactic relation depending of syntactic position. In (2a), *kúma* is a verb and it is a root, while in (2b), *kúma* is a noun, and it has an object relation. The problem is that we cannot apply the same logic to (5), as the tag VERB refers to a participle which can function as a verb and as an adjectives or noun.

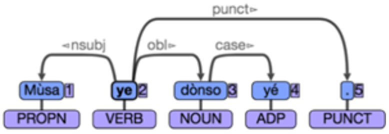
Another challenging issue was a part-of-speech tag for copulae. In Bambara, there are three main types of non-verbal predication: presentative, locative and equative. According to UD annotation scheme, functional words cannot be heads of lexical words. Auxiliaries, including copulae, are function words, they cannot have dependents which are expressed by nouns or adverbs. We have annotated copulae as VERB, as, if we change the aspect in the locative and equative constructions, copulae *bɛ́* and *yé* will be replaced by the verb *kɛ́* 'do' (6, 7).

(6a)



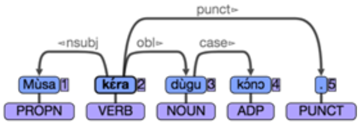
'Musa is in the village'

(7a)



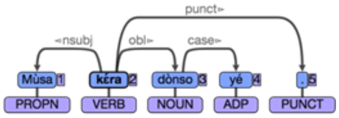
'Musa is a hunter'

(6b)



'Musa was in the village'.

(7b)

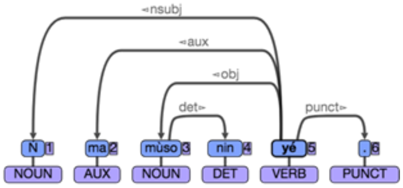


'Musa was a hunter'.

3.2.2 Double part-of-speech tags

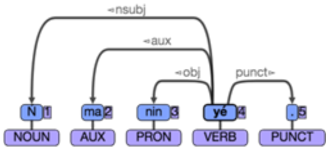
In the original BRC annotation scheme, some words were annotated with two part-of-speech tags. This was done in cases where a word could be annotated for part of speech differently according to syntactic context. For example, a word which could be a determiner or a pronoun would receive the tag dtm/prn (determiner or pronoun). In the (8a), *nin* is determiner, while in (8b), it is a pronoun.

(8a)



'I did not see this woman',

(8b)

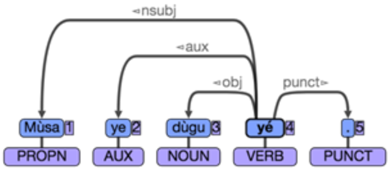


'I did not see this'.

3.3 Other conversion issues

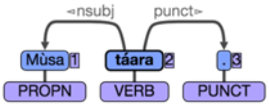
In CoNLL-U, the next column after the part-of-speech tag is that of grammatical features. While the list of part-of-speech tags is closed, the list of features is open, so one can add language specific features to the universal list. We used only one language specific feature for Bambara: Valency=1/2 for (in)transitive verbs. It was used earlier in the Ainu treebank (Senuma & Aizawa 2017). This feature is indispensable for Bambara, as there is a phenomenon of bidirectional case marking, which denotes an auxiliary that appears only in SOV structures, preventing linear adjacency of subject and object (Heath 2018; Nikitina 2018). This is the case with an auxiliary *yé* in (9a).

(9a)



'Musa saw a village'.

(9b)



'Musa left'.

Heads and syntactic relations are results of annotation process, so there was no conversion issues there. Enhanced dependencies are not currently available in annotation tool (cf. 4) which we were using. In the miscellaneous column, we kept glosses from BRC.

4 Annotation tool and guidelines

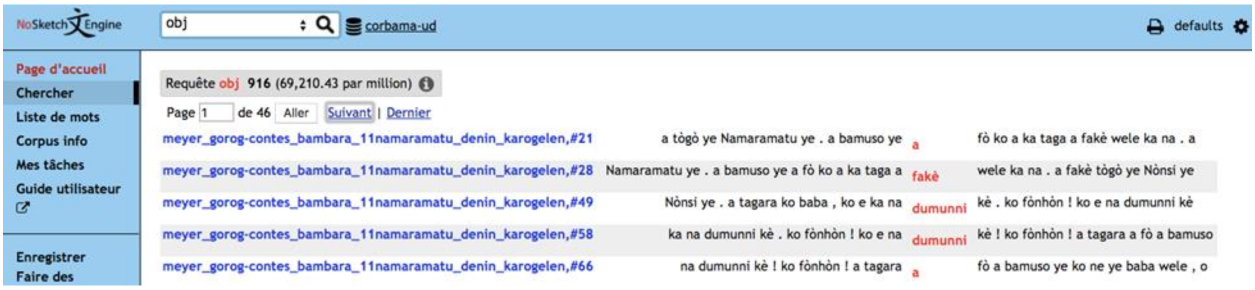
Sentences in CoNLL-U were annotated using UD Annotatrix[8] (Tayers & Sheyanova & Washington 2018). The main idea of this newly created annotator is to make dependency annotation faster, easier and more interactive in comparison with other existing tools. Sentences were annotated by me, then we checked them with Francis M. Tayers and Valentin Vydrin.

Annotation guidelines providing a description and examples for all syntactic relations used in the treebank are available on the UD site.[9]

5 Integration of a treebank into the BRC

A newly created Bambara treebank is currently available only on the site of BRC as corbama-ud subcorpus. When UD site is updated, new treebanks, including the Bambara one, will be available.

In the corbama-ud, one can use a simple query field for a search by a word or by a dependency relation. For example, in order to find all direct objects, one needs to indicate the *obj* relation. As a result, we get a concordance with all direct objects.



Requête obj 916 (69,210.43 par million)

meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#21	a tògò ye Namaramatu ye . a bamuso ye	a	fò ko a ka taga a fakè wele ka na . a
meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#28	Namaramatu ye . a bamuso ye a fò ko a ka taga a	fakè	wele ka na . a fakè tògò ye Nònsi ye
meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#49	Nònsi ye . a tagara ko baba , ko e ka na	dumunni	kè . ko fònhòn ! ko e na dumunni kè
meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#58	ka na dumunni kè . ko fònhòn ! ko e na	dumunni	kè ! ko fònhòn ! a tagara a fò a bamuso
meyer_gorog-contes_bambara_11namaramatu_denin_karogelen,#66	na dumunni kè ! ko fònhòn ! a tagara	a	fò a bamuso ye ko ne ye baba wele , o

Figure 2: Results of a simple query *obj*.

For more elaborate queries, one can use a CQL[10] field. In the corbama-ud, it is possible to search by following tags:

- WORD: original word form
- LEMMA: word in a standardized orthography
- TAG: UD part-of-speech tag
- FEATURE: UD feature.
- ID: index of a token Gramma
- HEAD: index of a head of a respective token
- DEP: UD syntactic relation
- GLOSS: glosses from BRC

[8] https://github.com/jonorthwash/ud-annotatrix
[9] http://universaldependencies.org/bm/dep/
[10] CQL stands for a Corpus Query Language, documentation is available here: https://www.sketchengine.eu/documentation/corpus-querying.

6 Summary and future plans

The paper gives an overview of the work carried out on the Bambara treebank. Together with Francis Tayers we developed an annotation scheme, then approximately 13 thousands tokens were annotated. This number is enough to train a parser, which is our immediate goal. When we get more annotated texts, we are planning to conduct quantitative researches about Bambara syntax. Moreover, there is a project of corpora for other Mande languages;[11] as soon as these corpora are built, it will be possible to create treebanks for these languages too, which would open good perspectives for comparative syntactic researches.

7 References

Aplonova, E. & Tayers, F. (2018). Towards a dependency-annotated treebank for Bambara. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, 22-24 January 2018*. Charles University, Czech Republic.

Heath, J. (2018). Mande-Songhay. In Proceedings of the 5th Conference on Mande Languages and Linguistics, 18-19 April 2018. LLACAN, France.

Maslinsky, K. (2014). Daba: a model and tools for Manding corpora. In *Proceedings of TALAf 2014: Traitement Automatique des Langues Africaines, 1 July 2014*. Marseille, France.

Nikitina, T. (2018). Bidirectional case marking in Wan. In Proceedings of the 5th Conference on Mande Languages and Linguistics, 18-19 April 2018. LLACAN, France.

Senuma, H. & Aizawa, A. (2017). Towards universal dependencies for Ainu. In *Proceedings of the NoDaLida 2017 Workshop on Universal Dependencies, 22 May 2017,* Gothenburg, Sweden.

Tayers, F., Sheyanova, M. & Washington, J. (2018). UD Annotatrix: An annotation tool for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, 22-24 January 2018*. Charles University, Czech Republic.

Vydrin, V. (2013). Bamana reference corpus. In *Procedia – Social and Behavioral Science*, 95: 75-80.

Vydrin, V. (2107). Bamana Jazyk (Bamana Language). In *Yazyki Mira: Yazyki Mande (Languages of the World: Mande Languages)*, pp.: 46-143, Institute of Linguistics, Russia.

[11] http://cormand.huma-num.fr/mandeica/