Towards a dependency-annotated treebank for Bambara

Ekaterina Aplonova
School of Linguistics
ВШЭ
Moscow
aplooon@gmail.com

Francis M. Tyers
School of Linguistics
ВШЭ
Moscow
ftyers@hse.ru

Abstract

In this paper we describe a dependency annotation scheme for Bambara, a Mande language spoken in Mali, which has few computational linguistic resources. The scheme is based on Universal Dependencies. We describe part-of-speech tags, morphological features and dependencies and how we performed a rule-based conversion of an existing part-of-speech annotated corpus of Bambara, which contains approximately 900,000 tokens. We also describe the annotation of a small treebank of 116 sample sentences, which were picked randomly.

1 Introduction

One of the basic language resources has, for a long time, been a part-of-speech tagged (or morphologically disambiguated) corpus. In recent years, treebanks — collections of sentences annotated for syntactic structure — have become increasingly available and vital resources, both for natural-language processing and corpus linguistics. Current end-to-end pipelines like UDPipe (Straka et al., 2016), which perform each stage of the classic NLP pipeline from tokenisation to dependency parsing, make it easy to go from a situation where a language has no effective language resources to one where the language has a functional pipeline in a few months as opposed to a few years of work.

A crucial prerequisite for building a treebank is to have a set of annotation guidelines which describe how particular syntactic structures are to be represented. In our work on creating a treebank for Bambara we have chosen version 2.0 of the Universal Dependencies scheme (Nivre et al., 2016) as it provides ready-made recommendations on which to base annotation guidelines for part-of-speech tags, morphological features and dependency relations. This reduces the amount of time needed to develop bespoke annotation guidelines for a given language as where the existing *universal* guidelines[1] are adequate they can be imported wholesale into the language-specific guidelines. In addition, the Universal Dependencies project provides a free/open *pool* (in the terminology of Streiter et al. (2006)) which collects dependency corpora in a single place, allowing for economies of scale in maintenance and ensuring that resources can persist after any initial development effort.

The remainder of the paper is laid out as follows: In Section 2, we give a short typological overview of Bambara, in Section 3, we describe an existing annotated resource for Bambara, the *Corpus Bambara de Référence* (CBR). Section 4 describes the conversion process we used, Section 5 describes some constructions in Bambara, which are not typologically common, and how we intend to annotate them. Finally, Sections 6 and 7 describe future work and conclusions respectively.

2 Bambara

In the description of Bambara presented in this paper, we used as sources the Vydrin (2013) and Выдрин (2017).[2] Bambara is the most widely-spoken language of the Manding language group (Western Mande < Mande < Niger-Congo). It is spoken mainly in Mali by 13–14 million people; of these, around four

[1] http://universaldependencies.org/guidelines.html

[2] Abbreviations are as follows: PFV = perfective predicative marker; SG = singular; POSS = possessive postposition; ; AG.OCC = suffix, which denotes an occasional actor; PP = postposition; QUAL = special predicative marker for qualitative verbs; PRES

Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16), pages 138–145,
Prague, Czech Republic, January 23–24, 2018.

Keywords: Bambara, Mande, treebank

Section	Sentences	Tokens
Unannotated	—	4,113,006
Disambiguated	—	903,585
Dependency annotated	116	1307

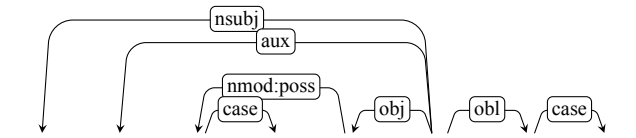
Table 1: Composition of the *Corpus Bambara de Référence* as of December, 2017

million are L1 speakers. There are two variants of naming this language: Bambara and Bamana, both of them are in use. Bambara is one of 13 "national languages" of Mali. Besides French, it is the major language on Malian radio and television, there are periodicals in Bambara, it is broadly used in literacy programmes and in primary schools; it is also taught at several universities in Europe and the US.

Bambara is a tonal language. It has two level tones and a down-drift. Tones can be lexical and grammatical, i.e. every lexeme has its lexical tone(s), which can change depending on the context into grammatical one(s). For instance, in the noun phrase *jíri fin* 'black tree', the rule of tonal compactness is demonstrated when the recessive syllables take the tone of dominant ones. The lexical tone of *fin* 'black' is low, however, in an attributive position, it takes the tone of its head *jíri* 'tree', whose tone is high. Moreover, in Bambara, there is a tonal definite article (indicated by a low floating tone). In the CBR it is not indicated. For this reason, in the present paper, we do not indicate it either. Tones are never marked in Bambara press and books published in Mali; tonal notation is present in publications of texts by linguists, however, even in the latter case it desperately lacks uniformity.

As described by Vydrin (2013), Bambara is an isolating language with certain elements of agglutination and incorporation. The basic word order is S AUX O V X. Therefore, in (1) *Fúla* is a subject, *ye* is an auxiliary, *á ká mìsi* is a direct object, *dí* is a verb, *í mà* is an oblique.

(1)



'The Fula man gave you his cow'.

The word order is fixed, however it is possible to remove a topicalised NP in the beginning of the clause (see §5.5).

3 Corpus Bambara de Référence

Development of the Bambara Reference Corpus (usually known by its name in French, *Corpus Bambara de Référence*) was started in April 2012. It is composed of texts of different kinds e.g. periodicals, oral literature, manuals, religious publications, letters from newspaper readers, texts recorded and transcribed by researchers etc. Since the Bambara orthographic standard is relatively undeveloped, the corpus assumes different levels of orthographic normalisation. The corpus includes a non-disambiguated sub-corpus and a disambiguated one (see Table 1 for statistics about its composition). In the non-disambiguated sub-corpus, there is only Bambara texts without any annotation. Annotation in the disambiguated sub-corpus, consists of part-of-speech tags, glosses and a respective token in a normalized orthography (with tones). A user is able either to search the entire corpus or to limit their search to the disambiguated sub-corpus. Texts have been and continue to be disambiguated by volunteers using Daba (Maslinsky, 2014), a morphological analyser based on a language-independent framework dictionary and

= presentative copula; LOC = locative copula; EQU = equative copula; NEG = negative copula; IPFV = imperfective predicative marker; INF = infinitive predicative marker; QUOT = quotative 'copula'; PROH = prohibitive predicative marker; PL = plural; NP = noun phrase; REL = relative pronoun/determinative; PFV.NEG = negative perfective predicatve marker; DIN = suffix, which derives a dynamic verb from a qualitative verb.

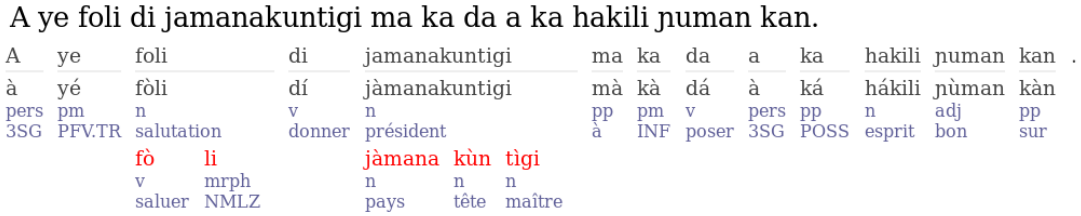


Figure 1: Example of a disambiguated sentence. The output format is machine-readable HTML. A free translation of the sentence into English would be 'He greeted the president and swore that he has good thoughts'. The first line is the original text, the second line is the tokenised text, the third line are the lemmas, and the fourth line has the part-of-speech tags. The fifth line has a gloss following the Leipzig glossing rules. Subsequent lines give a morphemic breakdown and gloss.

a rule-based morphological analyser. Language-specific data used by the analyser consist of a dictionary and a list of rules for splitting words into morphemes. Its output consists of seven lines: sentence in the original orthography, separate tokens in the original orthography, separate tokens in normalised orthography, part of speech tags, separate morphemes and glosses (cf. Figure 1).

4 Data conversion

To convert the corpus we used a Python script which reads the HTML format of the CBR performed substitution of tags and wrote the output in CoNLL-U format. In order to generate the morphological features, it was necessary to look at both the glosses and the morphological breakdown of the words.

We were able to maintain the original tokenisation scheme for the sentences, with the exception of three auxiliaries, the affirmative progressive marker *bɛ́ kà*, the negative progressive marker *tɛ́ kà*, and the emphatic perfective marker *yɛ́ kà*, which we treat as fixed units as it is not possible to give part-of-speech tags to the individual parts.[3]

Regarding the lemmas, we left them as in the original corpus, where they appear as word forms with the addition of tone marking. We do not treat compounding and derivation productively, so the lemma of the compound *jamanakuntigi* 'president' is not split into its component parts *jamana-kun-tigi* 'country-head-master'.

Part-of-speech tags were largely able to be converted deterministically using a simple translation table, however there was one tag, `conj` 'conjunction' which needed to be split into `CCONJ` 'co-ordinating conjunction' and `SCONJ` 'subordinating conjunction'. For this we made a list of lemmas for both types, and converted based on this.

In the original 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 pronoun would receive the tag `dtm/prn` (determiner or pronoun). The majority of determinatives perform different syntactic functions, e.g. the same word can act as an argument or as an attribute. Another example would be the tag `conj/prep` (conjunction or preposition). Prepositions are closely connected to some subordinate conjunctions. There are only seven prepositions and each of them can also act as a subordinate conjunction. These lexemes are treated as preposition, if they introduce a NP. If they introduce a whole clause, they are treated as conjunctions. We manually resolved these ambiguities, annotating them with the appropriate universal tag according to context.

We used the following language-specific features for Bambara: `AdjType=Attr` was used for adjectives with the suffix -/man/ and `Valency=1` was used for intransitive verbs, while `Valency=2` was used for transitive verbs. The feature `AdjType=Attr` is also used in the Afrikaans treebank to mark attributive adjectives (in Afrikaans adjectives have separate attributive and predicative forms). The feature `Valency=1` has been proposed for use in the Ainu treebank (Senuma and Aizawa, 2017).

[3] A reviewer suggests that we could have these as separate tokens with the part of speech tag `AUX` for both parts and the dependency relation `fixed`. As this would allow us to maintain the same tokenisation as the original we are planning to implement this change.

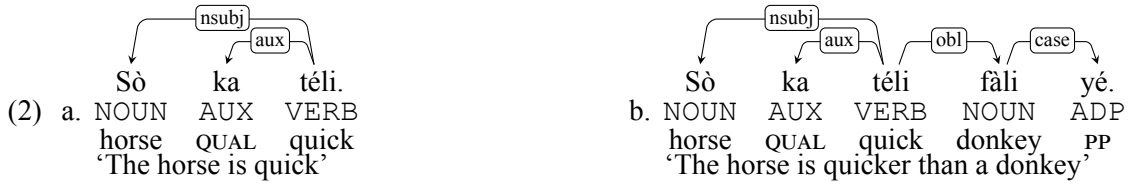
In addition to converting the part-of-speech tags and morphological features a number of sentences were annotated for dependencies using the UD ANNOTATRIX annotation tool (Tyers et al., 2018) by a single annotator in discussion with various linguists while developing the guidelines.

5 Dependency scheme

In this section, we describe some of the features of Bambara, which are not typologically common, and how they are annotated. We use the original glosses (partly modified in order to make it clearer for readers, who are not familiar with Bambara) along with dependency relations from Universal Dependencies.

5.1 Qualitative verbs and adjectives

In Bambara, verbs are divided into two classes: dynamic verbs and qualitative verbs. Qualitative verbs have special predicative marker, glossed as *q*ual. They cannot express tense, aspect, modality values (2a). Moreover, they cannot bear a direct object, but they can have adjuncts (2b). We annotated them as VERB.

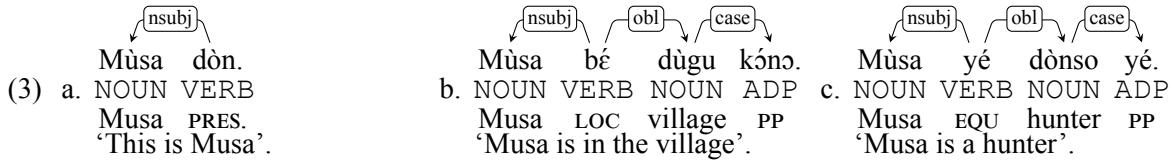


In predicative position, adjectives can be used only as secondary predicates. In the main predicative position, there are only qualitative verbs.

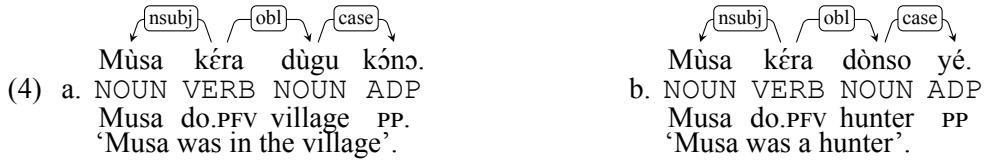
A considerable number of adjectives are derived from qualitative verbs by adding a suffix -/man/: téli 'quick' → téliman. However, there are two other types of adjectives, which do not have a suffix -/man/. In the first type, there are adjectives derived from qualitative verbs by conversion: mɔ́gɔ fìn 'black (adj) man (lit. 'black man')' → mɔ́gɔ ká fìn 'a man is black (verb)'. In the second type, there are simple (non-derived) adjectives: *kúra* 'new', *gánsan* 'simple', *sèbɛ* 'serious', *bèlebele* 'fat, bai', *bánga* 'without sauce', etc.

5.2 Non-verbal predication

There are three main types of non-verbal predication: presentative (3a), locative (3b) and equative (3c).

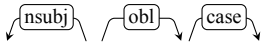


We annotated all copulae as VERB. First of all, in presentative construction, the copula *dòn* is always the last element of a clause. We cannot postulate an ellipsis of a predicate, so this is the copula, which bears all predicative functions. Secondly, if we change an aspect in locative and equative constructions, the copula will be replaced by a verb *kɛ́* 'do' (4a, 4b).



In negative clauses, in all these three types of predication, the negative copula *té* is used (5).

(5)



Mùsa tɛ́ dònso yé.
NOUN VERB NOUN ADP
Musa EQU hunter PP
'Musa is not a hunter'

5.3 Infinitive marker

A verbal infinitive form is unmarked morphologically. It is introduced by a predicative marker *kà*. Verbs introduced by *kà* cannot bear their own subjects, but they can bear objects and obliques.

An infinitive construction can be an argument of the verb in the main clause (6a), its adjunct with the purpose meaning (6b) and it can express a sequential meaning (6c). We annotated *kà* as AUX.

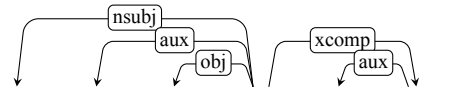
(6) a.

N bɛ sé kà móbili bòli.
PRON AUX VERB AUX NOUN VERB
1SG IPFV arrive INF car run
'I can drive'.

b.

Ù ká ɲógɔn sɔ̀rɔ kà bɛ̀nkan sɔ̀rɔ.
PRON AUX PART VERB AUX NOUN VERB
3PL SUBJ together find INF agreement find
'They met together in order to find an agreement'.

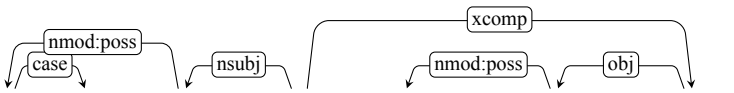
c.



Dúnan ye jí mìn kà kúma.
NOUN AUX NOUN VERB AUX VERB
guest PFV water drink INF speak
'A stranger drunk a water, (then) he began to speak'.

Note that verbs of motion *táa* 'go' and *nà* 'come' take a verbal complement phrase without infinitive marker (7).

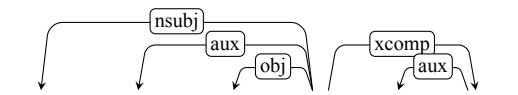
(7)



Dàa ká cí-den táa-ra Farabugu dùgu-tigi wéele.
NOUN ADP NOUN VERB NOUN NOUN VERB
Dah POSS send-child go-PFV Farabugu village-master call
'Dah's messager went to call the chief of Furabugu'.

The dependency relation is *xcomp*, because if a predicate of a main clause is negated, the subordinate clause is in the scope of negation (8).

(8)



Dúnan ma jí mìn kà kúma.
NOUN AUX NOUN VERB AUX VERB
guest PFV.NEG water drink INF speak
'A stranger drunk a water, (then) he did not speak'.

5.4 Quotative ‘copula’

In CBR, *kó* is always annotated as a *quotative copula*, however, Выдрин (2017) mentions, that, perhaps, we could postulate several homonymous lexemes. In (9a), the word *kó* has its own subject and it introduces direct speech, but in (9b), it only introduces a subordinate clause.

(9) a.

‘The girl said: I am going to the village’.

b.

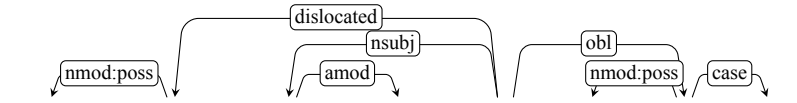
‘All tell him that he should not be afraid’

If *kó* has its own subject, we annotate it as `VERB`, unless it is annotated as `SCONJ`.

5.5 Topicalisation

Any NP can be placed in the beginning of the sentence and, thus, topicalised. A resumptive pronoun takes its place (10).

(10)



‘Residents of Bamako, many of them are there’

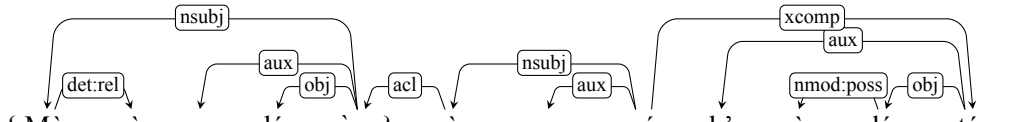
This strategy is commonly used for introducing the subject of a main clause. We annotate the topicalised NP as `dislocated` and the resumptive pronoun gets the main function of the NP.

5.6 Adnominal clauses

Adnominal clauses include relative clauses and participle clauses. There are two main relativisation strategies. In the first strategy a dependent clause precedes the main clause (11), while in the second one a subordinate clause follows the main clause (12).

In the first strategy, the two clauses are combined into what, from a functional point of view, is a relativising construction: one of the clauses narrows the potential reference of a referring expression from the other clause.

(11)

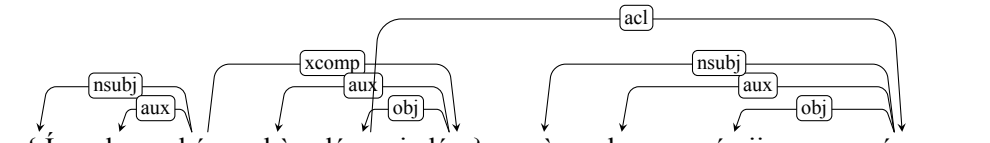


‘A woman who had a child cannot leave her child’.

In (11), a relativised noun is followed by a determinative *mìn* in the subordinative clause, while in the main clause there is a resumptive pronoun *ò*, which refers to this noun. As mentioned by Nikitina (2012), such a strategy does not fit into any of widely recognised relativisation strategies.

The second type (12) is more typologically common. It is a simple correlative strategy similar to European languages: a relativised noun in the main clause has a pronominal referent in the dependent clause.

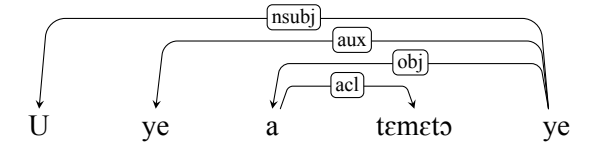
(12)



'You should eat a meal which will increase the quantity of breastmilk'.

There are also adnominal clauses which are not relative clauses (they are not marked with a relativiser). This goes for the participle forms -/len/, -/ta/, -/bali/ and for the converb -/tɔ/ (13).

(13)



'They have seen him passing by'.

These are also annotated with the `acl` relation.

6 Future work

In terms of linguistic analysis, there are a number of avenues for future research. Bambara syntax is understudied and we would like to work on our analysis of relativisation strategies, the quotative *kó* and the various predication/copula markers.

In terms of the treebank, the immediate objective is to annotate 10,000 tokens in order to solidify the annotate scheme and produce a first version. After this, we aim to annotate up to 100,000 tokens. We are planning to compile an annotation guide available to download. The work will be continued as part of the first author's masters thesis work. Moreover, there are also corpora for other Mande languages which could be annotated under a similar scheme and we would also like to experiment with cross-lingual parsing for this language group.

7 Concluding remarks

We have presented a large part-of-speech annotated corpus converted to Universal Dependencies along with a small proof-of-concept section annotated for dependency relations. We have described how a number of constructions in Bambara can be annotated and laid out the future work for the corpus.

Acknowledgements

Many thanks to Valentin Vydrin for his help and encouragement and for permission to use the Corpus Bambara du Référence. We thank the anonymous reviewers for their extremely helpful advice and comments.

References

Maslinsky, K. (2014). Daba: a model and tools for Manding corpora. In *Proceedings of TALAf 2014 : Traitement Automatique des Langues Africaines*, pages 114–122.

Nikitina, T. (2012). Clause-internal correlatives in Southeastern Mande: A case for the propagation of typological rara. *Lingua*, 122:319–334.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of Language Resources and Evaluation Conference (LREC'16)*.

Senuma, H. and Aizawa, A. (2017). Toward universal dependencies for Ainu. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 133–139.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France. European Language Resources Association (ELRA).

Streiter, O., Scannell, K., and Stuflesser, M. (2006). Implementing NLP projects for non-central languages: Instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4):267–289.

Tyers, F. M., Sheyanova, M., and Washington, J. N. (2018). UD Annotatrix: An annotation tool for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, page [this volume].

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

Выдрин, В. (2017). Бамана язык. In Vydrin, V., Mazurova, Y., Kibrik, A., and Markus, E., editors, *Языки мира: Языки манде*, pages 46–143. РАН. Институт языкознания.

A Supplemental material

Table 2 gives the conversion table for part-of-speech tags from the CBR to UD annotation schemes. The conversion table for morphological features is too long to include here but may be found online.[4]

Description	CBR	UD POS	UD Feats
Adjective	adj	ADJ	
Adverb	adv	ADV	
Postpositional adverb	adv.p	ADV	
Expressive adverb	adv.ex	ADV	
Numeral	num	NUM	
Noun	n	NOUN	
Proper noun	n.prop	PROPN	
Verb	v	VERB	
Qualitative verb	vq	VERB	
Participle	ptcp	VERB	VerbForm=Part
Personal pronoun	pers	PRON	PronType=Prs
Pronoun	prn	PRON	
Modal word	pm	AUX	
Copula	cop	VERB	
Conjunction	conj	CCONJ SCONJ	
Postposition	pp	ADP	
Determiner	dtm	DET	
Particle	prt	PART	

Table 2: Conversion table for the parts of speech. Choice of conjunction type is determined lexically.

[4] https://github.com/KatyaAplonova/UD_Bambara