A morpheme-based treebank for Gbaya, an Ubanguian language of Central Africa

Paulette Roulon-Doko
CNRS, LLACAN, France
`paulette.roulon@cnrs.fr`

Sylvain Kahane
Paris Nanterre University,
CNRS, MoDyCo, France
Institut Universitaire de France
`sylvain@kahane.fr`

Bruno Guillaume
Universite de Lorraine,
CNRS, Inria, LORIA, France
`Bruno.Guillaume@inria.fr`

Abstract

In this paper, we present the first treebank for Gbaya, a language from the under-resourced Niger-Congo family. The language has a rich system of tonal morphemes and virtually no affixes. The dependency analysis is based on a morpheme-based tokenisation and the treebank is also distributed in a word-based Universal Dependencies version. Several constructions are discussed in the paper: genitive construction, clause coordination, sentence particles, adverbial and relative clauses, serial verb constructions, reported speech, topicalization, and focalization.

1 Introduction

This paper presents the first treebank for a Gbaya language. We have decided to follow the tokenization proposed by Roulon-Doko (1995). In this previous work, the analysis considers many tones as autonomous units that combine with root lexical tonal pattern for grammatical reasons; there are called grammatical tones (Hyman, 2016; Rolle, 2018). These tones are placed after the determined term. In the case of verbs without lexical tonal pattern, the tone is attached to the verb base and is therefore always indicated beforehand. In order to encode this tokenization level, we use the mSUD framework (Guillaume et al., 2024). mSUD was designed for morpheme-based level annotation; it is a variant of the Surface Syntactic Universal Dependencies (SUD) (Gerdes et al., 2018a) annotation schema.[1] In this paper, we will used the term "morpheme" also for units that represent tones. We will consider three kind os tokens: roots, clitics, and inflectional morphemes (which are mainly tonal tokens).

The corpus is also available[2] in the Universal Dependencies (UD) framework (de Marneffe et al., 2021). To meet the UD tokenisation requirements, we provide automatic conversion to the UD format, in which tones are not expressed as separate tokens.

The corpus is made up of three tales. The tales in Gbaya are a repertoire without specialists. The language used to tell tales is the language of everyday life, with no stylistic form of its own. The storyteller, whether male or female, young or old, takes the floor spontaneously during a storytelling session. These three tales were recorded in 1970 in the village of Ndongué (Central African Republic) during traditional storytelling evenings by Paulette Roulon. Tale T16-C6, which tells the story of the woman fishing at the dam whose baby was swept away by the water after the dam burst, opened the session. It was told by Anna Zàngé, a woman in her forties. This was followed by tale T9-C7, which tells the story of the brother and sister who went on a hunting camp, which was then told by Yvonne Yàì-sɔ̀ a young girl of around 17. Tale T24-C59, which tells the story of Wanto and the little cob, was told at another session by Hélène Dúî, a young woman in her thirties. The Table 1 gives an overview of the length of each tale in the treebank **mSUD_Northwest_Gbaya-Autogramm**.

Northwest Gbaya (ISO: GYA, WALS: gbk, Glottocode: `nort2775`) is part of the main linguistic group Gbaya-Mandja-Ngbaka, an Ubangian language family (a branch of the Niger-Congo phylum, Adamawa Ubangi). It is spoken in the northwest of the Central African Republic (CAR) and in the central-eastern part of Cameroon. It is subdivided in six dialects: four in CAR [ɓòdòè, ɓòkpàn, ɓòɲìnà, ɓùgùì], usually named Gbaya-kara, and two in Cameroon [ɓòya, yàáyùwèè]. In

---

[1] The mSUD corpus is available at `https://github.com/surfacesyntacticud/mSUD_Northwest_Gbaya-Autogramm`.

[2] `https://github.com/UniversalDependencies/UD_Northwest_Gbaya-Autogramm`

	audio length	# of sentences	# tokens (morph-based)	# tokens (word-based)
T16-C6	321s	143	1,363	847
T9-C7	178s	79	820	503
T24-C59	359s	181	1,621	1,067
TOTAL	**858s**	**403**	**3,804**	**2,417**

Table 1: Sizes of the 3 samples of the corpus **mSUD_Northwest_Gbaya-Autogramm** (version 2.15).

1996, there were 265,000 speakers: 200,000 in the western part of CAR and 65,000 in the central-eastern part of Cameroon. This paper deals with the Gbáyá ɓòdòè, a Northwest Gbaya dialect spoken in Central African Republic. The annotation is based on a dictionary from Roulon-Doko (2008) and a complete grammar, based on a 4h50 oral corpus of spontaneous speaking, collected in the field between 1970 and 2013, processed with Toolbox and Elan, glossed and translated into French, by Roulon-Doko.

The consonant system has three glottalized consonants (ɓ, ɗ, ʔ), four labio-velars (kp, gb, ngb, ɲm), oral consonants (both voiceless and voiced) and a complete range of nasals and semi-nasals. There are seven oral and five nasal vowels. Gbaya is a tonal language with two levels and four tones (H, L, LH, HL). All vowels carry a tone. Gbaya syllabic structure includes open and closed syllables, but no initial vowel.

Gbaya is an isolating language with very little morphology and no agreement at all. Gbaya relies minimally on derivation but makes strong use of compounding, marked in writing by a hyphen between components (e.g. *gèέ-fìò* ceremony sp.), which is also used for adjectives-adverbs with a reduplicated structure (e.g. *bàɗàm-bàɗàm* irregularly arranged).

The lexemes are thus simple, compound, or structurally reduplicated. They are distributed across 18 categories (Roulon-Doko, 2008). Gbaya has verbs, non-verbal predicates, nouns and four subcategories of adjectives. The main lexical categories, VERB (10%), NOUN (50%), and ADJ (32.6%), have unrelated lexical stocks. For nouns, composition is very important (47.6%) and derivation very little used (3.7%).

Word order is very strict. Gbaya is an SVO language, that also makes use of non-verbal predication. The subject is compulsory, except for some construction that will be described below.

In Section 2, we explain how the tone system is encoded as "morphemes" in the mSUD framework. The Section 3 describes the principles used for defining the maximal units. Several intersting contructions of Gbaya are described in Section 4, together with the proposed analysis in the treebank. The last Section 5 is dedicated to the conversion to the UD treebank.

2 The morpheme-based annotation

Verb inflection in Gbaya is marked by tone alternations, possibly accompanied by affixes. Verbal lexemes have no lexical tonal pattern, the tonal pattern they carry systematically comes from the TAM marker and is identical for all. The TAM verbal system is organized in three moods (Realis, Virtual and Command) and two aspects (Perfective and Imperfective). Tense is not marked on the verb in Gbaya (Roulon, 1975; Roulon-Doko, 1994). Verbs always express a process where the obligatory subject is either external to the process (transitive construction = transitive voice) or included in the process (intransitive construction = middle voice).
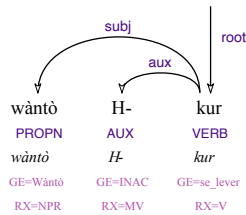
2.1 Tonal tokens

Tonal tokens, or morphotonemes (Meeussen, 1967), are inflectional morphemes that impose tonal patterns or trigger tonal alternations. Three cases of tonal tokens are discussed.
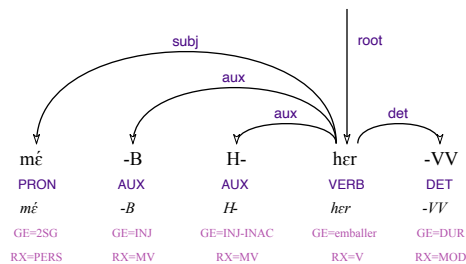
TAM markers. The 19 TAM markers are all tonal tokens. They include a verbal pattern that we place in front of the verb root (R) for the sake of regularity, even if they contain suffixal elements. For TAM markers, we follow the UD annotation scheme: they receive **upos=AUX** and depend on the root by a **aux** relation: see sentences (1) and (2).[3]

---

[3]In fact, TAM marker are distributional heads, because they control the distribution of the verb forms, which is different between finite and non-finite forms. In consequence, auxiliaries are generally treated as heads in SUD. Due to the particular status of Gbaya's TAM markers, which are only inflectional morphemes, we have treated them as dependents, following UD.

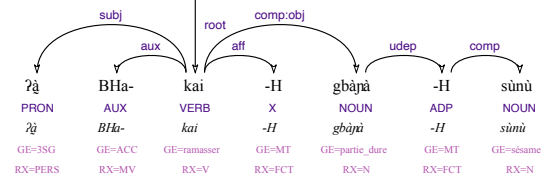(1) *Wàntò kúr*
   H-kur
   NPR IPFV-get_up
   'Wanto gets up.'

subj		root
	aux	
wàntò	H-	kur
PROPN	AUX	VERB
wàntò	*H-*	*kur*
GE=Wàntò	GE=INAC	GE=se_lever
RX=NPR	RX=MV	RX=V

(2) *mὲ hɛ́r*
   mɛ́-B H-hɛr-VV
   2SG-INJ INJ_IPFV-wrap-DUR
   'Wrap it.'

subj			root	
	aux			det
		aux		
mɛ́	-B	H-	hɛr	-VV
PRON	AUX	AUX	VERB	DET
mɛ́	*-B*	*H-*	*hɛr*	*-VV*
GE=2SG	GE=INJ	GE=INJ-INAC	GE=emballer	GE=DUR
RX=PERS	RX=MV	RX=MV	RX=V	RX=MOD

(3) *ʔá̰ kàyá gbàɲá sùnù*
   BHa-kai-H gbàɲá-H
   3SG PFV-gather-MT hard_part-MT sesame
   'He collects sesame seeds.'

subj		root	comp:obj		udep	comp
	aux		aff			
ʔà̰	BHa-	kai	-H	gbàɲà	-H	sùnù
PRON	AUX	VERB	X	NOUN	ADP	NOUN
ʔà̰	*BHa-*	*kai*	*-H*	*gbàɲà*	*-H*	*sùnù*
GE=3SG	GE=ACC	GE=ramasser	GE=MT	GE=partie_dure	GE=MT	GE=sésame
RX=PERS	RX=MV	RX=V	RX=FCT	RX=N	RX=FCT	RX=N

(4) *dáà kɔ́ bêm*
   father of child
   'The childs father.'

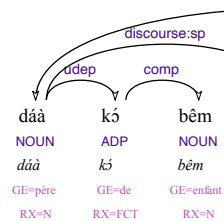|		
	discourse:sp	
	udep	comp
dáà	kɔ́	bêm
NOUN	ADP	NOUN
dáà	*kɔ́*	*bêm*
GE=père	GE=de	GE=enfant
RX=N	RX=FCT	RX=N

In (1), we show the realis imperfective (**IPFV**), which is realized by a high tone H-. In (2), we show the injunctive-imperfective (**INJ_IPFV**), which is also realized by a high tone H- on the verb, but also imposes a low tone -B on its subject, which must obligatorily be a personal pronoun. Such a construction forces us to make choices when converting the treebank to the word-based UD annotation scheme. The fusion of the personal pronoun and the low tone could be analyzed an alternative form of the pronominal subject (our choice) or with an auxiliary bearing a pronominal index. If the second case is common in many languages, it would have been uncommon in Gbaya, which has no other cases of pronominal indices and where clauses always have an overt subject.

The floating tone -H. There are regular tonal modifications linked to the presence of a floating high tone -H, which occurs between two words. It modifies the tonal pattern of the first word depending on the first tone of the second word according to fixed rules, except in the case of a high pattern. This floating high tone can mark grammatical distinctions (functioning as a supra-segmental marker, TM in the glosses). It occurs in two con-
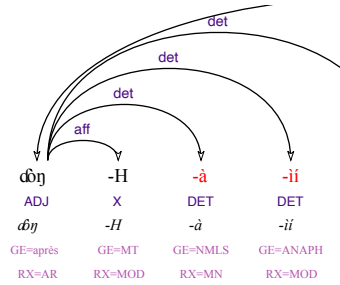
texts. Either (i) it is structurally attached to certain words (preposition, perfective verbal marker, etc.), involving no choice on the part of the speaker, similar to class markers, for instance (as the first -H in (3)); or (ii) it functions as a connective linking the two nouns of a genitive phrase (second -H in (3)) contrasting with the genitival SN with segmental connective *kɔ́* (4).
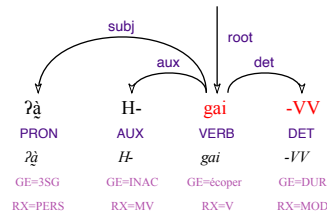
In the first case, it is analyzed as an affixe of first word, with **upos=X** and linked to the root by a relation **aff** (it represents 75% of the occurrences in the treebank). In the second case, it is analyzed with **upos=ADP**, like the real **ADP** *kɔ́*. In SUD, **ADP**s are analyzed as the head of the adpositional phrase and the noun depends on the **ADP** by a relation **comp**.

Multicategory markers. Muticategory markers are suffixes or morphotonemes that can combine with lexemes of different parts of speech. We treat all of them as suffixes and give them **upos=DET**. They are the anaphoric *-ìí*, the locative *-ὲ*, the insistent -V polar tone, the durative -VV, the suffix *-à* (definite, nominalizer). They are not exclusive from each other. See (5) and (6).
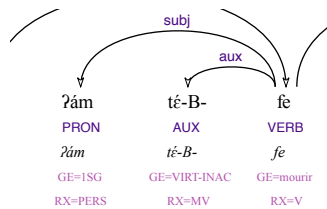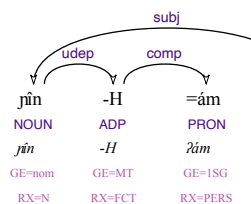
(5) *dòɲáìí*
dɔ̀ŋ-H-á-ìí
behind-MT-NMLS-ANAPH
'After that'

dɔ̀ŋ -H -à -ìí
ADJ X DET DET
dɔ̀ŋ *-H* *-à* *-ìí*
GE=après GE=MT GE=NMLS GE=ANAPH
RX=AR RX=MOD RX=MN RX=MOD

(det, det, det, aff)

(6) *ʔà gáííí*
ʔà H-gai-VV
3SG IPFV-bail-DUR
'She bails (the water) for a long time'

ʔà H- gai -VV
PRON AUX VERB DET
ʔà *H-* *gai* *-VV*
GE=3SG GE=INAC GE=écoper GE=DUR
RX=PERS RX=MV RX=V RX=MOD

(subj, aux, root, det)

(7) *ʔám té-fè*
ʔám té-B-fe
1SG VIRT-IPFV-die
'I'm going to die.'

ʔám té-B- fe
PRON AUX VERB
ʔám *té-B-* *fe*
GE=1SG GE=VIRT-INAC GE=mourir
RX=PERS RX=MV RX=V

(subj, aux)

(8) *ɲínám*
ɲîm-H-ám
name-MT-1SG
'My name'

ɲîn -H =ám
NOUN ADP PRON
ɲîn *-H* *ʔám*
GE=nom GE=MT GE=1SG
RX=N RX=FCT RX=PERS

(subj, udep, comp)

(9) *bám*
H-ba-ám
INJ_IPFV-catch-1SG
'Catch me'

H- ba =m
AUX VERB PRON
H- *ba* *ʔám*
GE=INJ-INAC GE=prendre GE=1SG
RX=MV RX=V RX=PERS

(ubj, aux, aux, root, mod, comp:obj)
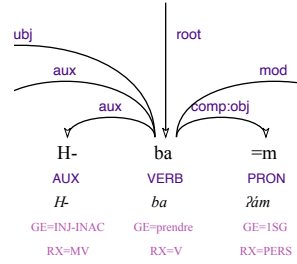
2.2 Clitics

Personal pronouns have a free form when they are used as a subject and placed before the verb (cf. *ʔám* in (7)). When they are placed after any term, they cliticized on it and loose the initial consonant when it is the glottal /ʔ/ (cf. *=ám* in (8) and (9))

All personal pronouns are treated in this way, even those that begin with a different consonant and have the same form in all positions (2SG *mɛ́* vs *=mɛ́*, 3PL *wà* vs *=wà*)

3 Sentence segmentation

Every verb has an overt subject, except in two constructions, relative clauses and coordination of clauses: two successive clauses that share the same subject can be coordinated without repeating the subject before the verb in the second clause. See for example sentence (10).

This gives us a very simple criteria to segment our spoken corpus into sentences: a sentence contains one and only one non-subordinated verb with a subject. A coordinated clause with the main clause will be attached the same sentence if and only if the verb has no subject. This follows the prescriptions of previous spoken corpora (Kahane et al., 2021).

Additionally, there's a case of close coordination without an explicit marker: when two clauses share the same subject, the subject is omitted in the second clause. This subject omission, which goes against the norm of repeating the subject, is annotated with conj:coord, linking the second verb (without subject) to the first.

(10) *wà   dɛ́   nɛ̀   sádàkà   kɔ́ɡà   ʔèá   hɔ́ɡɔ́ɔ́ɔ́   yɔ́   nɛ̀   mɔ̀*
3PL IPFV-do CIRC celebration of-3SG only for_a_long_time IPFV-dance CIRC thing
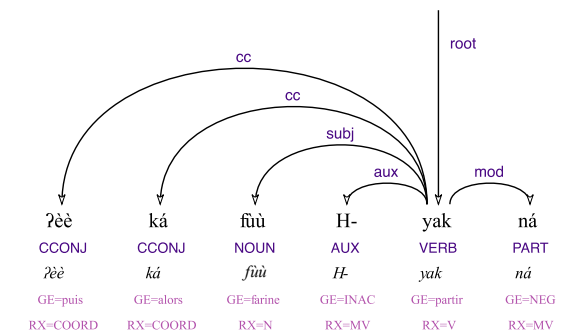*kádí̧*
wholly

'They just celebrate her for a long time and they dance until they're exhausted.'
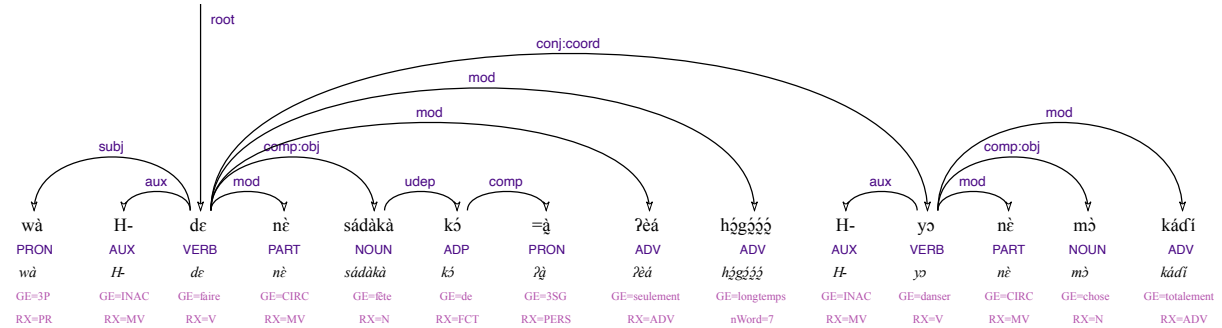
(11) *ʔàm   kpáwà*
ʔB-àm
and_so-1SG IPFV-find-3PL

'And so I met them'
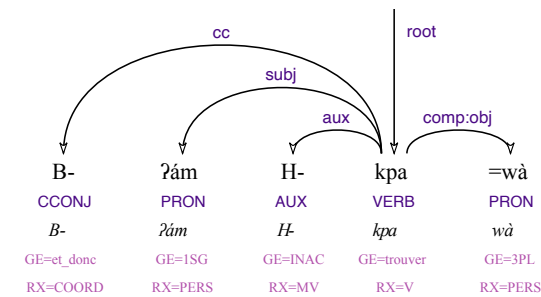
(12) *ʔèè   ká   fùù   yák   ná*
coord coord
then   and_then flour IPFV-leave NEG

'Then the flour didnt leave again'

4 Some syntactic constructions

We give an overview of several syntactic constructions, some of them unusual in the current UD collection, such as the reported speech construction.

4.1 Sentence coordinating conjunction

Clauses are usually coordinated, expressing the speaker's enunciative choice. There are eight coordinating conjunctions (plus four variants), such as *ʔá̧-nɛ̀* 'and now', *ká* 'and then' or B- 'and so' a low tone on the subject personal pronoun (11). The coordinating conjunction *ʔèè* 'then' can combine with all the others coordinating conjunctions (12).

4.2 Sentence particles

In addition to the coordinating conjunctions, Gbaya has a system of nine enunciative particles, which are always placed at the end of a clause and serve an enunciative function by specifying the speakers point of view. We categorize them

as **PART** and introduce the relation **discourse:sp** (**sp** for sentence particle), already used for Chinese treebanks (Leung et al., 2016).

The sentence particles must be distinguished from the discourse markers, which are analyzed as interjections (**INTJ**) and receive the function **discourse** (without an extension).

4.3 Adverbial clauses

In SUD, subordinating conjunctions are analyzed as the head of the clause, with the verb as a **comp** (13).

In Gbaya, however, many subordinate clause relationships are not marked by a subordinating element; rather, it is the verb form itself that marks the subordination. These are the TAM-bound forms. We provide the example of the hypothetical, which renders the clause dependent on a main

(13) kàà   pí   ʔɛ́ráà   kɔ̀ gúé   ká   gúé   ká   gúé   ʔá́   hɛ̀ɛ́-hɛ̀
      when-3SG   IPFV-put   hand-MT-3SG   in fish_hole and_then fish_hole and_then fish_hole BE-LOC   just_here

'When she puts her hand in the fish hole, it's a fish hole all right.'
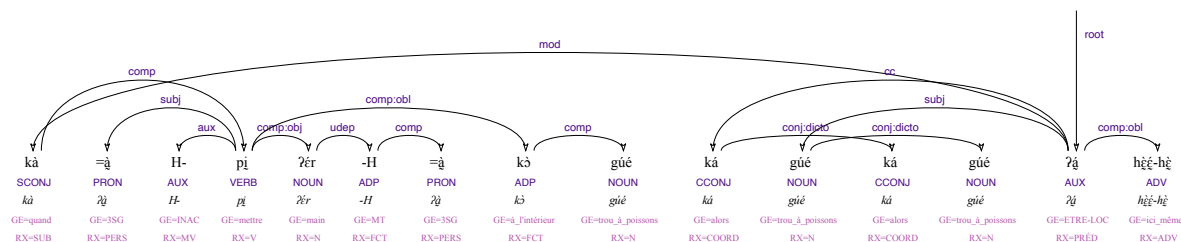
clause that must follow (see example (14) where the subordinative clause is in blue).

4.4 Relative clauses

There is no relative pronoun in Gbaya, but only one pure relativizer *nɛ̀*, which is analyzed as a subordinating conjunction. In 11 of the 18 examples of our treebank, the subject is extracted and the clause has a gap in the subject position (see (15) where the relative clause is in blue and the domain name in orange).

Object relative clauses have also a gap in the object position, but in locative relative clause there is an adverb in the extracted position. Gbaya has also an original construction with a nominalized verb as the antecedent, which must be repeated in the relative clause (16).

4.5 Serial verb construction

Serial verb constructions consists of the expansion of a verb in predicative position by another verb in the infinitive (perfective, imperfective, or virtual), within the limit of a sequence of three verbs V1.TAM V2.INF [V3.INF]. In the treebank, V1 functions as the root, V2 is linked to V1 by a **compound:svc** relation (17).

4.6 The reported speech

Reported speech constitutes an original construction in Gbaya, combining two clauses and two enunciations. These two clauses are interdependent neither subordinated, as in subordinative clauses, nor sequential, as in coordinated clauses. The first clause, referred to as the "quoting speech", introduces the speaker; the second, the "quoted speech', provides the content of the discourse. Without going into the specifics of this construction, it is important to note that the quoted speech alone is sufficient to establish reported discourse. The quoting speech is rarely verbal and

is often limited to the speakers name (eventually followed by a particle, the most common being *ndé*). The quoted speech may begin with a mention of the addressee, which can be followed by the particle *nà*. Note that *ndé* have the primary role of being a sentence particle expressing interrogation, and *nà* have the primary role of being a verb modifier expressing negation. We analyze both the speaker and the addressee with a new relation which we call **discourse:participant**. See (18) and (19).
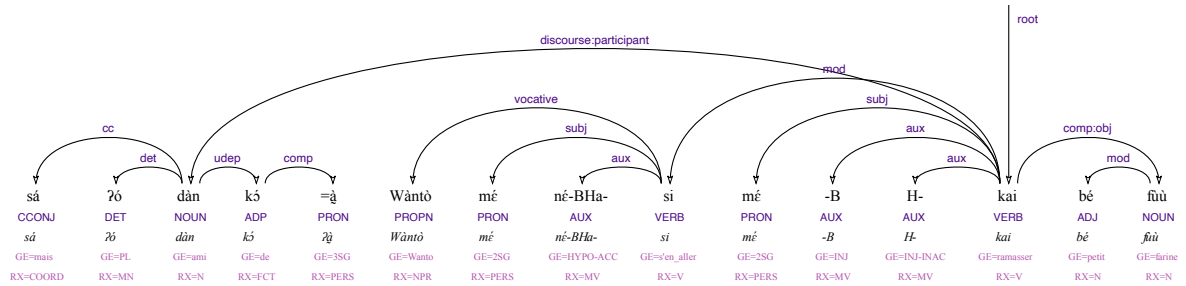
4.7 Topicalization

In Gbaya, topicalization brings a noun or noun phrase (the topic) to the front of the sentence, followed without pause by the comment. When the subject occupies this initial position, specific markers (a resumptive pronoun, a topicalization particle, or both) are required to indicate its topical status. Moving a direct object or oblique complement to the front of the sentence is enough to topicalize it, sometimes reinforced by a topicalization marker. If the object is animate, a resumptive pronoun follows the verb; if it is inanimate, the position remains empty. Another construction topicalizes the verb by placing a verbal noun first, followed immediately by the same verb.

4.8 Focalization

In Gbaya, noun focalization is marked by the identifiers *nɛ́* or *mɛ̀-nɛ́ / mɔ̀-nɛ́*. The sentence begins with this identifier, followed by a clause in which the focused element functions as subject, object, or circumstantial complement. Such a construction exhibits a cleft extraction that splits the verbal predication into two parts. Only *nɛ́* combines with the potential verb form to express negation, unlike *mɛ̀-nɛ́*, which occurs only in affirmative clauses. Verbal focus uses *nɛ́*, functioning like an applicative suffix placed after the conjugated verb, target-

(14)

Wàntò	mɛ́	nɛ́-sìá	mɛ̀	káí	bé	fùù
NPR	2SG	HYPO-PFV-return	2SG-INJ	INJ_IPFV-collect	little	flour

'Wanto if you return, take some flour.'

Dependency tree (14):

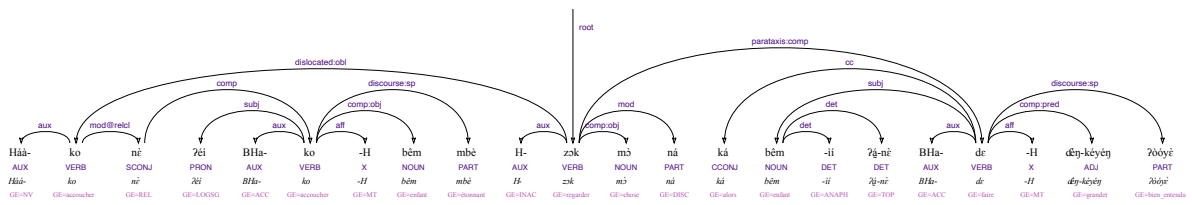| form	sá	ʔó	dàn	kɔ́	=à̀	Wàntò	mɛ́	nɛ́-BHa-	si	mɛ́	-B	H-	kai	bé	fùù
POS	CCONJ	DET	NOUN	ADP	PRON	PROPN	PRON	AUX	VERB	PRON	AUX	AUX	VERB	ADJ	NOUN
	sá	ʔó	dàn	kɔ́	ʔà̀	Wàntò	mɛ́	nɛ́-BHa-	si	mɛ́	-B	H-	kai	bé	fùù
GE	mais	PL	ami	de	3SG	Wanto	2SG	HYPO-ACC	s'en_aller	2SG	INJ	INJ-INAC	ramasser	petit	farine
RX	COORD	MN	N	FCT	PERS	NPR	PERS	MV	V	PERS	MV	MV	V	N	N

(15)

ká	nɔ́	yérè	nè	gbɛ̀ɛ̀	bókáɲáà	ɲɛ́m	ɓɔ̀	ná
and_then	fat	buffalo	REL	PFV-kill-MT	sister-MT-3SG	IPFV-suit	yet	NEG

'The way buffalo fat dulled his sisters senses is just unbelievable.'

Dependency tree (15):

form	ká	nɔ́	-H	yérè	nè	BHa-	gbɛ	-H	bókáɲá	-H	=à	H-	ɲɛm	ɓɔ̀	ná
POS	CCONJ	NOUN	ADP	NOUN	SCONJ	AUX	VERB	X	NOUN	ADP	PRON	AUX	VERB	ADV	PART
	ká	nɔ́	-H	yérè	nè	BHa-	gbɛ	-H	bókáɲá	-H	ʔà̀	H-	ɲɛm	ɓɔ̀	ná
GE	alors	graisse	MT	buffle	REL	ACC	tuer	MT	soeur	MT	3SG	INAC	suffire	encore	NEG
RX	SUB	N	FCT	N	SUB	MV	V	FCT	N	FCT	MN	MV	V	ADV	MV
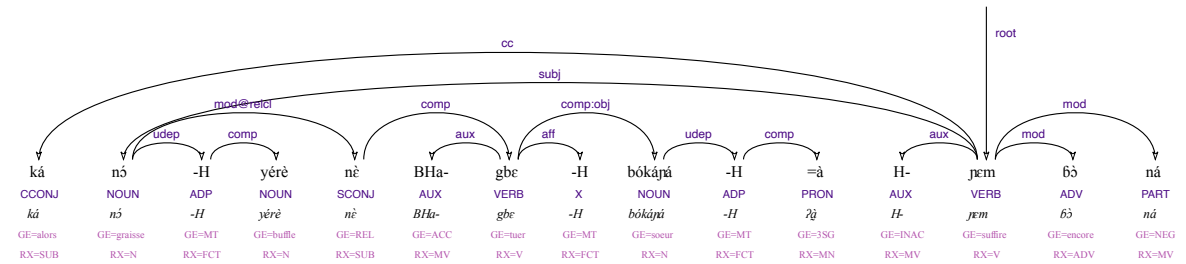
(16)

kóáà	nè	ʔéí	kòò	bêm	mbè	zɔ̀k	mɔ̀	ná	ká
NV-give_birth	REL	LOGSG	PFV-give_birth-MT	child	amazing	IPFV-see	thing	DISC	and_then

bêmí	ʔá-nè	dɛ̀ɛ̀	dɛ́ɲ-kéyéɲ	ʔòóyɛ̀
child-ANAPH	TOP	PFV-dMT	tall_(6_8_years_old)	of_course

'Having surprisingly given birth to a child, she realizes that the child is of course grown up [6-8 years].'

Dependency tree (16):

form	Háà-	ko	nè	ʔéí	BHa-	ko	-H	bêm	mbè	H-	zɔk	mɔ̀	ná	ká	bêm	-íí	ʔá-nè	BHa-	dɛ	-H	dɛ́ɲ-kéyéɲ	ʔòóyɛ̀
POS	AUX	VERB	SCONJ	PRON	AUX	VERB	X	NOUN	PART	AUX	VERB	NOUN	PART	CCONJ	NOUN	DET	DET	AUX	VERB	X	ADJ	PART
	Háà-	ko	nè	ʔéí	BHa-	ko	-H	bêm	mbè	H-	zɔk	mɔ̀	ná	ká	bêm	-íí	ʔá-nè	BHa-	dɛ	-H	dɛ́ɲ-kéyéɲ	ʔòóyɛ̀
GE	NV	accoucher	REL	LOGSG	ACC	accoucher	MT	enfant	étonnant	INAC	regarder	chose	DISC	alors	enfant	ANAPH	TOP	ACC	faire	MT	grandir	bien_entendu
RX	MV	V	SUB	PERS	MV	V	MOD	N	MOD	MV	V	N	DISC	COORD	N	MOD	MOD	MV	V	MOD	AA	MOD

(17) Ɂéí   nɛ̀ɛ̀   kpéé   yì
             V1   V2
     LOGSG   PFV-go-MT   INF_IPFV-close   river

'I've gone to block the river (scoop fishing).'
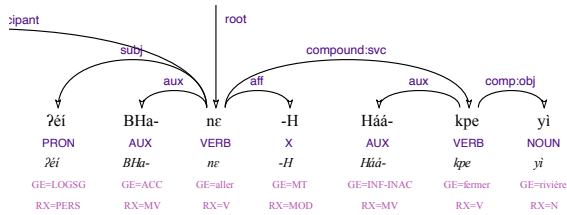
(18) Ɂá̰-nɛ̀   wà   wàntò   mɛ́   tɛ́-Ɂɔ̀
     and_now   they   Wanto   you   lying_down
     dîî
     where

'and now they (say): Wanto, where do you want to sleep?'
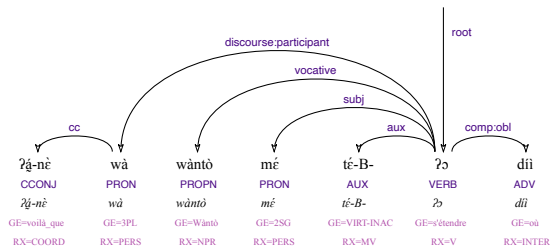
ing the immediately following element: the object (with transitive verbs), a circumstantial complement (with intransitives), or, if no complement follows, a verbal noun derived from the same verb.

5 UD version of the treebank

Like for the mSUD annotation, the UD version of the Gbaya treebank is produced in two steps. First, the word tokenisation is modified to match the syntactic word level expected by UD. Then the regular conversion for SUD to UD (Gerdes et al., 2018b) is used to produce the UD version.

5.1 Word-level tokenisation

Word-level tokenisation is obtained from the original annotation by merging inflectional morphemes with the tokens to which they are attached.

There are few cases where the inflectional morpheme is not directly syntactically related to its neighbour token. In the example below, the suffix -B is not syntactically linked to the previous token Ɂɛ́nɛ́. We have built a dedicated heuristic to compute the necessary syntactic structure after the fusion: in the example, we want to keep the **subj**

relation in the final structure.
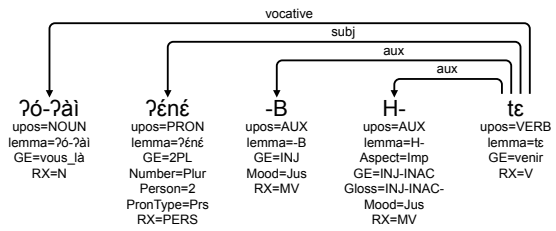
After affixes merging, the example above is converted into:
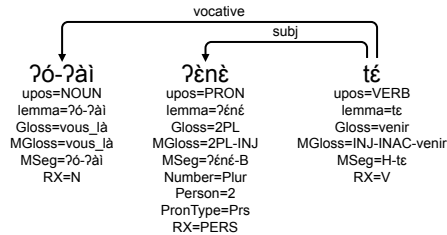
Note that in the word-based version, the morpheme-based analysis is indicated in the **MSeg** and **MGloss** features.

5.2 Conversion to UD

The annotation obtained in the previous step is not a regular SUD format. Nevertheless, the universal conversion from SUD to UD can be applied because the Gbaya word-level format is between SUD and UD: adpositions are heads, as in SUD, and auxiliaries are dependent, as in UD. Thus, only a subset of the conversion rules is activated during the conversion process. With sentence (19) the 3 annotation formats are given: first the annotated format at morpheme level, then the SUD annotation at word level and then UD annotation (also at word level).

6 Conclusion

The Gbaya treebank is the first treebank of a new genus within the Niger-Congo family, and only the Fourteenth treebank (along with Tswana, Yoruba and Wolof) of the greater Niger-Congo family, which comprises over 1,500 languages, including the 400 Bantu languages. Gbaya differs from Bantu languages in the absence of nominal classes and agreement rules. It has a highly developed tonal inflection system, but virtually no affixes. We have chosen to develop a morpheme-based treebank in order to highlight the elegance of this language's grammatical system and to have a resource that combines the interlinear glosses previously developed by Paulette Roulon-Doko with

(19) dáà   kɛ́ bêm ná   ʔéí   nɛ̀ɛ̀   kpéé   yì
     father of child DISC LOGSG PFV-go-MT INF_PFV-close   river

'[She says] To the childs father I go to block the river.'
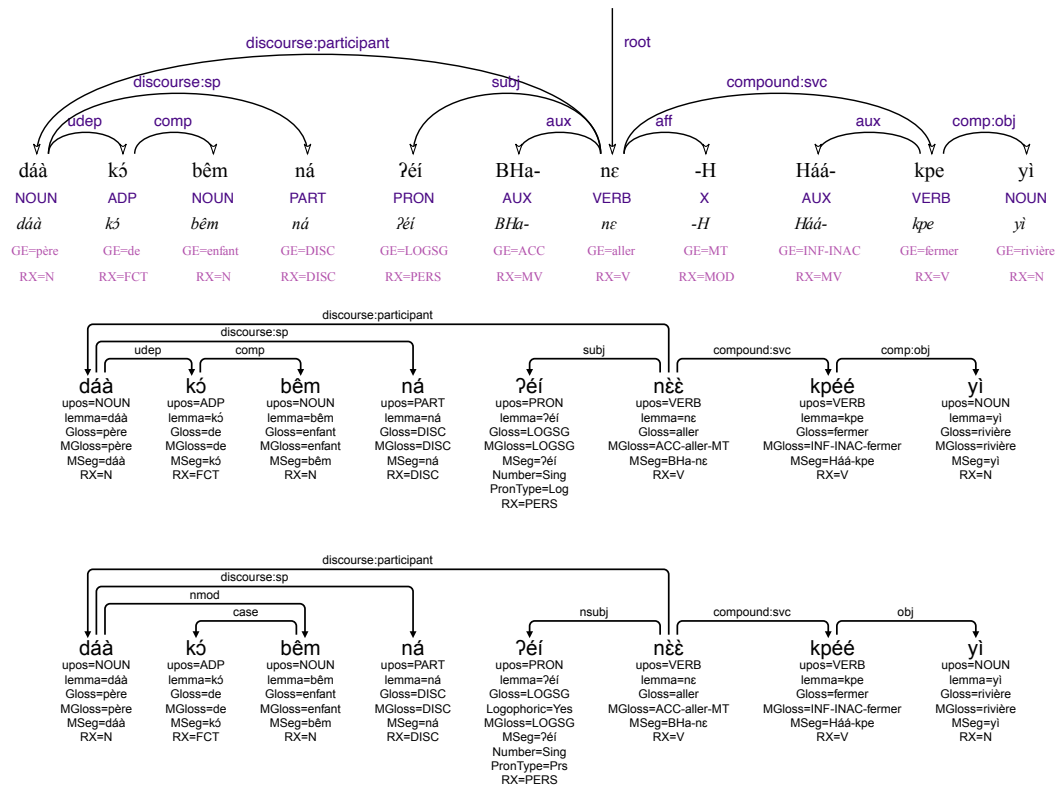
syntactic analysis. The treebank is also distributed in a word-based UD version, where morphological information remains accessible in the **MSeg** and **MGloss** features.

Acknowledgements

References

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018a. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Universal Dependencies Workshop 2018*, Brussels, Belgium.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018b. SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Bruno Guillaume, Kim Gerdes, Kirian Guiller, Sylvain Kahane, and Yixuan Li. 2024. Joint annotation of morphology and syntax in dependency treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9568–9577, Torino, Italia. ELRA and ICCL.

Larry M. Hyman. 2016. Lexical vs. grammatical tone: Sorting out the differences. In *Proceedings of Tonal Aspects of Language (TAL) 2016*, pages 5–11, Buffalo, New York. ISCA Archive.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Herman Leung, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes, and John Lee. 2016. Developing Universal Dependencies for Mandarin Chinese. In *Proceedings of the 12th Workshop on Asian

Language Resources (ALR12), pages 20–29, Osaka, Japan. The COLING 2016 Organizing Committee.

A. E. Meeussen. 1967. Bantu grammatical reconstructions. *Africana Linguistica*, 3(1):79–121.

Nicholas R. Rolle. 2018. Grammatical tone: Typology and theory. *Berkeley Papers in Formal Linguistics*, 1(1).

Paulette Roulon. 1975. *Le verbe gbaya, étude syntaxique et sémantique (R.C.A.)*. Number 51-52 in Bibliothèque de la SELAF. Paris. 2 cartes.

Paulette Roulon-Doko. 1994. L'expression de la qualification (l'exemple du gbaya 'bodoe de centrafrique). In T. Geider and R. Kastenholz, editors, *Sprachen und Sprachzeugnisse in Afrika*, pages 345–356. Rüdiger Köppe Verlag, Köln.

Paulette Roulon-Doko. 1995. Le système verbal gbaya. In R. Boyd, editor, *Le système verbal dans les langues oubanguiennes*, number 07 in LINCOM Studies in African Linguistics, pages 25–80. LINCOM, München.

Paulette Roulon-Doko. 1998. La prédication non processive en gbaya "bodoe. In P. Roulon-Doko, editor, *Les manières d'ń être ż et les mots pour le dire dans les langues d'Afrique Centrale*, pages 111–134. Lincom Europa, München-Newcastle.

Paulette Roulon-Doko. 2008. *Dictionnaire Gbaya-Français (République Centrafricaine), suivi d'un dictionnaire des noms propres et d'un index français-gbaya*. Coll. Dictionnaires et langues. Karthala, Paris.