Developing Universal Dependencies for Wolof

Cheikh Bamba Dione
University of Bergen / Sydnesplassen 7, 5007 Bergen
dione.bambal@uib.no

Abstract

This paper presents work on the creation of a Universal Dependency (UD) treebank for Wolof as the first UD treebank within the Northern Atlantic branch of the Niger-Congo languages. The paper reports on various issues related to word segmentation for tokenization and the mapping of PoS tags, morphological features and dependency relations to existing conventions for annotating Wolof. It also outlines some specific constructions as a starting point for discussing several more general UD annotation guidelines, in particular for noun class marking, deixis encoding, and focus marking.

1 Introduction

Wolof (ISO code: 693-3) is a Niger-Congo language mainly spoken in Senegal and Gambia.[1] Until recently, not many natural language processing (NLP) tools or resources were available for this language. Dione (2012a) developed a finite-state morphological analyzer. Dione (2014) reported on the creation of a deep computational grammar for Wolof based on the Lexical Functional Grammar (LFG) framework. That grammar has been used to create the first treebank for this language, making an important contribution to the development of the LFG parallel treebank (Sulger et al., 2013).

Treebanks play an increasingly important role in computational and arguably also theoretical linguistics. A treebank can be defined as a collection of sentences that typically contain various kinds of morphological and syntactic annotations (Abeillé, 2003). In recent years, different language processing applications (e.g. question answering, machine translation, information extraction) require high-quality parsers. Reliable and robust parsing models can be trained and induced from treebanks (Manning and Schütze, 1999).

The basic assumption in dependency grammar is that syntactic structure consists of lexical elements linked by binary asymmetrical relations called *dependencies* (Tesnière, 1959). The arguments to these relations consist of a head and a dependent. The head word of a constituent is the central organizing word of that constituent. The remaining words in the constituent are considered to be dependents of their head. Figure 1 shows an example of dependency structure from the WTB for the sentence[2] given in (1).

(1) *Noonu laa mujj a tànn beneen mecce, jàng dawal awiyoŋ.*
ADV 1SG.NSFOC finally.do to choose another profession learn pilot airplane
'So then I chose another profession, and learned to pilot airplanes.'

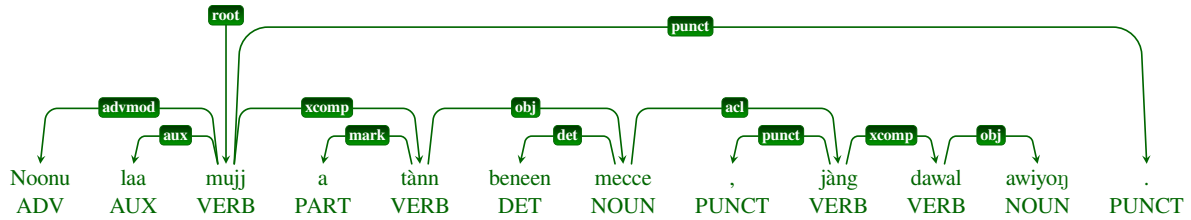


Figure 1: Example of a dependency structure from the WTB

[1] See http://www.ethnologue.com/language/WOL.
[2] Source: Wolof translations of *The Little prince* (Saint-Exupéry, 1971) available from http://www.wolof-online.com.

This paper presents work on the development of a Universal Dependency (UD) treebank for Wolof (henceforth WTB). It is the first effort in building dependency structures for Wolof in particular, and for the Northern Atlantic branch of the Niger-Congo languages in general. The annotations contained in the Wolof LFG treebank (henceforth WolGramBank) served as a basis for the creation of a scheme for the WTB. Note, however, that the WTB is not an automatic conversion from the LFG treebank, but was rather created manually (from scratch). This is mainly because such an automatic conversion (which is planed as future work) involves non-trivial mapping issues between LFG and UD. One of the most significant challenges is to determine which syntactic level of representation – constituency structure or functional structure – is the most natural basis for constructing dependency representations. Other crucial issues include e.g. the procedure of selecting the true head of syntactic constituents, the mapping from LFG to UD relations, the treatment of copula, coordination and punctuation (Meurer, 2017; Przepiórkowski and Patejuk, 2019).

The paper is structured as follows. Section 2 gives a brief overview of some salient features of the Wolof language. Section 3 describes the data collection process and the composition of the corpus. Section 4 discusses issues of word segmentation for tokenization. Section 5 describes the annotation processes for parts of speech (PoS), morphological features and syntactic relations. Section 6 concludes the discussion.

2 Background on Wolof

Before we take up the issue of the creation of a treebank for Wolof, we need to provide the reader with a general understanding of some salient features of that language.

2.1 Nouns, noun classes and determiners

Like the other Atlantic languages, Wolof has a noun class (NC) system (Greenberg, 1963; Sapir, 1971; McLaughlin, 1997) that consists of approximatively 13 noun classes:[3] 8 singular, 2 plural, 2 locative, and 1 manner noun classes. Like in Bantu languages, the Wolof noun class system also encodes *Number*. However, class membership is not marked on the noun itself, but rather on the noun dependents like determiners (e.g. articles, demonstratives), but also on (indefinite, interrogative and relative) pronouns and adverbs (locatives, manner). The noun classes are identified by their index (*b*, *g*, *j*, *k*, *l*, *m*, *s*, *w* for singular NCs and *y*, and *ñ* for plural NCs). The index "appears in the form of a single consonant on nominal dependents such as determiners and relative particles" (McLaughlin, 1997, p. 2).

Wolof determiners agree in noun class with the head noun. Determiners for different noun classes are distinguished by a consonant that is final (i.e. as a suffix) in the indefinite article (2c) and word-initial (i.e. as a prefix) in all other determiners. In addition, definite determiners encode information about proximity and distance with respect to the noun reference. As shown in (2), the definite article is constructed by suffixing a spatial deictic, *-i* for the proximal (2a) or *-a* for the distal (2b), to the consonantal class marker.[4]

(2) a. *xaj b-i*
dog NC-DFP
'the dog (proximal)'

b. *xaj b-a*
dog NC-DFD
'the dog (distal)'

c. *a-b xaj*
INDF-NC dog
'a dog'

Wolof has a rich system of demonstratives (Robert, 2016). These combine indications of the distance and reference point with respect to the speaker or addressee. For instance, for the *b* noun class, the four most commonly used forms are (*bii*, *bale*, *boobale*, and *boobu*), as exemplified in (3) with the noun *xaj* "dog".

(3) a. *xaj bii* 'this dog' (close to me, wherever you may be)
b. *xaj bale* 'that dog' (far away from me, wherever you may be)
c. *xaj boobale* 'that dog' (far away from both of us, but closer to you than to me)
d. *xaj boobu* 'that dog' (close to you and far away from me)

[3] The number of noun classes may vary according to dialects (Tamba et al., 2012).

[4] Abbreviations in the glosses: ADV: adverb; COP: copula; DEM: demonstrative; DET: determiner; DFP: definite proximal; DFD: definite distal; GEN: genitive; INDF: indefinite; LOC: locative; IPFV: imperfective; NC: noun class; NSFOC: non-subject focus; OBJ: object; POSS: possessive; PRES: present; PROG: progressive; PST: past tense; PL: plural; SG: singular; SFOC: subject focus; SUBJ: subject; VFOC: verb focus; 1, 2, 3: first, second, third person.

In Wolof, noun class membership is determined by a number of factors, including phonological, semantic and morphological criteria (McLaughlin, 1997; Tamba et al., 2012). For instance, many nouns that begin with [*w*] are in the *w*-class. Concerning morphology, nouns derived with certain derivational suffixes (e.g. *-in*) are assigned a specific class (e.g. the *w*-class). Finally, regarding semantics, trees typically are in the *g*-class, while most fruits are in the *b*-class. Also, the singular human noun class is the *k*-class, while the default plural human noun class is the *ñ*-class. However, the aforementioned factors just point to few tendencies found in the language. In fact, for each class, there are several words that do not follow these factors. The Wolof noun class system lacks semantic coherence (McLaughlin, 1997). The same can be said for the phonological and the morphological criteria. None of these factors are systematic indicators of noun classes in Wolof.

Furthermore, Wolof nouns are typically not inflected except for the genitive and the possessive case. Wolof genitives (4) are head-initial and show affinities with the Semitic construct state (Kihm, 2000). Such constructions involve a possessed entity described as the head and a possessor as its complement. The genitive relationship is overtly marked on the head noun by means of the *-u* suffix (e.g. *kër-u*) which precedes its complement (*buur* "king"). This suffix may also appear in other constructions like (5), which, unlike (4), do not denote possession, but rather seems to be just a normal compound, despite the similarity between these two constructions. In many other compounds like (6), the genitive marker does not appear at all.

(4) *kër-u buur*
house-GEN king
'king's house'

(5) *ndox-u taw*
water-GEN rain
'rain water'

(6) *téere xam-xam*
book knowledge
'knowledge book'

2.2 Adjectives

Wolof has no category for adjectives (Church, 1981; McLaughlin, 2004). The 'adjectival' concepts in Indo-European languages are typically expressed by stative verbs in Wolof. Adjectival constructions are realized as relative clause structures with the "adjective" being inflected like verbs.

2.3 Verbal system

In Wolof, a verb constituent has two components (Robert, 1991; Robert, 2000). The first component is the verb which is typically an invariant (unless derived) lexical stem. The second component is an inflectional marker that conveys the grammatical specifications of the verb, including person, number, tense, aspect, and mood features as well as the information structure of the sentence (focus). The inflectional marker can be preposed, postposed, or suffixed to the lexical stem, resulting in ten different paradigms or conjugations (Robert, 2010). Among these paradigms, we can distinguish non-focused conjugations from focused ones. Non-focus conjugations include perfective (7-8) and imperfective (9) constructions.

(7) *Xaj b-i lekk na.*
dog NC-DFP eat 3SG
'The dog has eaten.'

(8) *Lekk na.*
eat 3SG
'She/he/it has eaten.'

(9) *Xaj b-i di-na lekk.*
dog NC-DFP IPFV-3SG eat
'The dog will eat.'

Like Arabic (Attia, 2007) and many other languages, Wolof is a pro-language. This means that the subject can be explicitly stated as an NP or implicitly understood as a pro-drop. The pro-drop nature of the language is illustrated in the affirmative perfective examples given in (7-8). While (7) has an explicit subject, (8) does not. Nevertheless both sentences are grammatical. In (8), there is no overt subject, because the language freely allows the omission of such an argument. In examples (7-8), *na* is an agreement marker. It carries information about number, and person, which enables the reconstruction of the missing subject in (8).

Wolof has three focus conjugations: subject focus, verb focus, and complement focus. As these names imply, these constructions vary according to the syntactic function of the focused constituent: subject, verb, or complement. The latter has a wide meaning and refers in general to any constituent which is neither subject nor main verb. Table 1 illustrates the inflections for the verb *lekk* 'to eat' and the object *jën* 'fish' in the three focus types. As can be seen, focus is marked morphosyntactically.

The examples (10), (11) and (12) illustrate subject, verb and non-subject focus constructions, respectively.

	Subject focus	Verb focus	Complement focus
1SG	*maa* lekk jën	*dama* lekk jën	jën *laa* lekk
2	*yaa* lekk jën	*danga* lekk jën	jën *nga* lekk
3	*moo* lekk jën	*dafa* lekk jën	jën *la* lekk
1PL	*noo* lekk jën	*danu* lekk jën	jën *lanu* lekk
2	*yeena* lekk jën	*dangeen* lekk jën	jën *ngeen* lekk
3	*ñoo* lekk jën	*dañu* lekk jën	jën *lañu* lekk

Table 1: Subject, verb and complement focus in Wolof.

(10) *Faatu moo lekk jën.*
Faatu 3SG.SFOC eat fish
'It's Faatu who ate fish.'

(11) *Faatu dafa lekk jën.*
Faatu 3SG.VFOC eat fish
'What Faatu did is eat fish.'

(12) *Jën la Faatu lekk.*
fish 3SG.NSFOC Faatu eat
'It's fish that Faatu ate.'

Morphologically, one can reconstruct the origins of the subject, verb and non-subject focus markers as *-a*, *da-* and *la-*, respectively. An evidence for such a reconstruction can be seen in examples where the focus marker amalgamates with a noun or a proper name, as shown in (13a). Here, the form *Faatoo* is a phonological contraction and can be decomposed in *Faatu* + *a*, as illustrated in (13b). The main difference between (10) and (13a) is that in the former the constituent Faatu is dislocated, while in the latter that constituent bears the subject function. Indeed, (10) could be translated as "Faatu, it's her who ate the fish".

(13) a. *Faatoo lekk jën.*
Faatu.SFOC eat fish
'It's Faatu who ate fish.'

b. *Faatu a lekk jën.*
Faatu SFOC eat fish
'It's Faatu who ate fish.'

3 Data collection

The basis for the development of the WTB is a corpus of natural text data selected from the following sources: OSAD,[5] Wolof Online,[6] Wolof Wikipedia,[7] and Xibaaryi.com.[8] Table 2 lists the sources of the corpora used for creating the Wolof UD treebank.

Source	Genres	# Docs	# Tokens	# Sentences
OSAD	didactic, expository	6	6269	265
Wolof Online	informative, narrative	18	12988	673
Wolof Wikipedia	encyclopedic	12	9232	500
Xibaaryi	informative	17	15095	669

Table 2: Texts and genres in WTB.

The selection of texts for the WTB was meant to satisfy the following criteria. First, the data should be freely available as far as possible. Second, the text types should be chosen which are interesting to typical UD users. The data selected from Wikipedia is freely available under a Creative Commons license, facilitating its annotation and distribution. Also, users interested in computational linguistics, corpus linguistics and language typology may prefer texts which resemble other treebank texts or are even available in other languages, such as Wikipedia. Third, a range of different genres should be covered. Accordingly, we include texts from other sources than Wikipedia. For those sources, it was necessary to first clarify copyright issues.

4 Tokenization and word segmentation

Syntactic analysis in UD is based on a lexicalist view of syntax (i.e. dependency relations hold between words). According to De Marneffe (2014), practical computational models gain from this approach. Following this, the basic units of annotation are syntactic (not phonological or orthographic) words. Therefore, clitics attached to orthographic words need to be systematically segmented for proper syntactic analysis.

[5] http://www.osad-sn.com
[6] http://www.wolof-online.com
[7] https://wo.wikipedia.org
[8] http://www.xibaaryi.com

Word segmentation for tokenization in Wolof is a non-trivial task due to an extensive use of cliticization (Dione, 2017). As in Arabic (Attia, 2007), function words such as prepositions, conjunctions, auxiliaries and determiners can attach to other function or content words. Like Amharic (Seyoum et al., 2018), clitics in Wolof may undergo phonological changes. They may assimilate with word stems and with each other, making it difficult to recognize and handle them properly. The phonological change is also exhibited in the written form where clitics are attached to their host. For proper segmentation, then, we need to recover the underlying form first. For example, the word *cib* 'in a', can be segmented into the preposition *ci* 'in' and the indefinite article *ab* 'a'. However, if we simply segment the first characters *ci*, the remaining form, *b* will not have meaning. Furthermore, a non-trivial issue is ambiguity of clitics. For instance, a form like *beek* can be split into *bi* 'the' and *ak* where *ak* can actuallly be interpreted as a conjunction 'and' or a preposition 'with'.

Table 3 provides examples of full form words consisting of stems with clitics. The first row of the table is to be read as follows: the preposition *ak* 'with' may encliticize to the verbal stem *daje* 'meet', yielding the surface form *dajeek*.[9] The other surface forms involve different grammatical categories (determiners, conjunctions, pronouns, auxiliaries, etc.) and occur in a similar manner.

Stem PoS	***Clitic PoS***	***Example***	***Word form***	***Literal translation***
VERB	PREP	*daje* 'meet' + *ak* 'with'	*dajeek*	'meet with'
	DET	*joxe* 'give' + ay 'some'	*joxeey*	'give some'
DET	PREP	*ba* 'the' + *ak* 'with'	*baak*	'the with'
	CONJ	*bi* 'the' + *ak* 'and'	*beek*	'the and'
PREP	DET	*ci* 'in' + *ab* 'a'	*cib*	'in a'
	PREP	*ca* 'about' + *ak* 'with'	*caak*	'about with'
NOUN	CONJ	*ndox* 'water' + *ak* 'and'	*ndoxak*	'water and'
NAME	CONJ	*Ali* 'Ali' + *ak* 'and'	*Aleek*	'Ali and ...'
ADV	PRON	*fu* 'where' + *nga* 'you'	*foo*	'where you ...'
PRON	AUX	*ko* 'him/her' + *di*	*koy*	'him/her' + IPFV
	AUX	*mu* 3SG + *a* SFOC + *di* IPFV	*mooy*	3SG SFOC + IPFV
CONJ	AUX	*te* 'and' + *di* IPFV	*tey*	'and' + IPFV
	DET	*mbaa* 'or' + *ay* 'some'	*mbaay*	'or some'

Table 3: Examples of cliticization in Wolof

A crucial segmentation issue concerns the focus markers discussed in section 2.3. In accordance with the UD guidelines, we split the focus markers into a pronoun and a focus morpheme. Thus, contracted forms like third singular subject focus marker *moo* were decomposed into *mu* (3SG) and *a* (subject focus marker). The same applies for *dafa* which becomes *da* (verb focus marker) + *fa* (3SG), though *fa* is an irregular form. In contrast, *la* does not combine with a pronoun. The direct consequence of splitting focus elements like *moo* is that, as shown in (14b), the proper noun *Faatu* occurs in a dislocated position before the clause, and is resumed within the clause by the co-referential pronoun *mu*, the subject of the verb *lekk* 'eat'.

(14) a. *Faatu moo lekk jën.*
Faatu 3SG.SFOC eat fish
'It's Faatu who ate fish.'

b. *Faatu mu a lekk jën.*
Faatu 3SG SFOC eat fish
'Faatu, it's her who ate fish.'

Tokenization and word segmentation were done semi-automatically using the Wolof finite-state tokenizer (Dione, 2017). This tool includes a clitic transducer that can detect and demarcate contracted morphemes, handling these as separate words. For some cases, a manual revision was necessary.

5 Annotation

There are a number of existing interfaces in use that allow for manual annotation of UD treebanks. These include BRAT (Stenetorp et al., 2012), Arborator (Gerdes, 2013) and Tred.[10] In this work, manual annotation was done using UD Annotatrix (Tyers et al., 2018). Unlike the aforementioned tools, UD Annotatrix is designed specifically for Universal Dependencies. It can be used in online and in fully-offline mode. The tool is freely-available under the GNU GPL licence.

[9]The long vowel [*ee*] in *dajeek* results from a coalescence of the final vowel of *daje* with the stem-initial vowel of the PREP *ak*.
[10]https://ufal.mff.cuni.cz/tred/

5.1 Parts of speech annotation

The PoS tag set used in the UD scheme is based on the Universal PoS tag set (Petrov et al., 2012) and contains 17 tags. Because we wanted to use existing PoS tag annotation for Wolof as starting point, a mapping between the tagset in the Wolof LFG and the UD PoS tagset was necessary. At the coarse-grained level, the Wolof LFG tag set contains 24 tags. Thus, the conversion of the parts of speech information in LFG treebank to the UD PoS tag set required some considerations. Since UD does not allow sub-typing of PoS tags or language-specific tags, we adhere to this restriction. Below we discuss issues in adapting the UD annotation scheme to the existing Wolof tagset.

5.1.1 Nouns

WolGramBank makes a distinction between proper nouns and other noun types. One main reason for this is that proper nouns generally do not appear with determiners (while common nouns and indefinite pronouns for instance do). This distinction starts at early preprocessing steps (during tokenization and morphological analysis). The functional information about the syntactic type as a proper noun and the semantic type as a name are respectively provided by the morphological tags *+PropNoun* and *+PropTypeName*. Proper nouns are assigned the *NAME* tag, making the mapping to the corresponding UD tag *PROPN* straightforward.

Concerning the other noun types, WolGramBank distinguishes three categories: *NOUN*, *NGEN* and *NPOSS*. The first category includes nouns without any inflection (e.g. *kër* "house"). The second and third categories refer to nouns inflected in the genitive (e.g. *kër-u* "house of") or in the possessive case (e.g. *kër-am* "his/her house"), respectively (see section 2.1).

In the WTB, all the three categories (common nouns, nouns inflected for genitive and those inflected for possessive) are mapped into the PoS category *Noun*. In terms of syntactic annotation, nouns with an apparent genitive marker are assigned the *nmod*[11] relation and are treated differently from those which do not show such an inflection, e.g. *téere* 'book' in (6). Nouns in the latter category are marked as *compound*. Using the UD features (FEATS), it was possible to further categorize the different forms, e.g. *Case=Gen* for the genitive and *Poss=Yes* for the possessive.

5.1.2 Determiners

In the WTB, determiners and quantifiers are assigned the *DET* category. A distinction between these categories can be made using features, e.g. *NumType=Card* for quantifiers, as it is done in some UD treebanks.

5.1.3 Adverbs

WolGramBank distinguishes between various types of adverbs, depending on whether an adverb modifies a verb, a clause, or introduces negation (e.g. negative particles). In the WTB, however, we define ADV for any kind of adverbs, and use the *Polarity* and *PronType* features (e.g. for relative/interrogative adverbs) to describe the type of adverb where necessary (the morphological features are discussed in section 5.2).

5.1.4 Verbs and auxiliaries

As discussed in section 2.3, Wolof verbs typically do not themselves carry inflectional markers. Instead, inflection is in many cases carried by so called inflectional elements that appear as separate words. The inflectional markers express a bunch of subject-related and clause-related features, including subject agreement, but also tense-aspect mood (TAM), polarity, and the focus in the sentence.

In the Wolof LFG Grammar, the inflectional markers are grouped under the category *INFL*. This category subdivides into four subcategories corresponding to the information whether the marker expresses subject focus, non-subject focus, verb focus and progressive. The *AUX* (for auxiliaries) tag is used mainly for the *di* imperfective marker (including its past tense inflected forms, e.g. *doon*). Furthermore, the tag *COP* is used for copula verbs and inflectional markers found in predicative constructions. This choice was motivated by the idea to provide a uniform analysis for both simple copula and clefts in Wolof, as both instantiate the same forms (Dione, 2012b).[12]

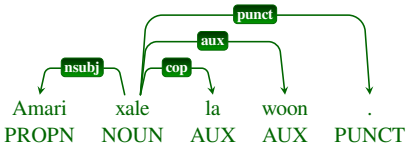
[11] *nmod* is used for nominal dependents of another noun and functionally corresponds to an attribute, or genitive complement.

[12] A more detailed discussion of the parallel syntactic proposed for copular and cleft clauses can be found in Dione (2012b).

However, the UD tagset scheme contains no *INFL* or *COP* tag. Still, it provides a general definition that allows for grouping these tags under the *AUX* category. UD defines an auxiliary as a function word that expresses grammatical distinctions not carried by the lexical verb, such as person, number, tense, mood, aspect. This is also the category provided for nonverbal *TAME* markers found in many languages. Thus, this is the category that fits the *INFL* tag from the Wolof LFG grammar. However, to keep the relevant information regarding the encoded information structure and copulae, it was necessary to introduce a new feature called *FocusType*. Such a feature is used to distinguish auxiliaries marking focus from other auxiliaries.

The UD guidelines state that the AUX category also includes copulas (in the narrow sense of pure linking words for nonverbal predication). Following this, the *COP* category from the LFG treebank was mapped to *AUX* in the Wolof UD treebank. This mapping, however, raised a small issue: in the UD scheme *AUX* cannot have a dependent, while in the existing annotation scheme for Wolof it is sometimes necessary for *COP* to have a dependent. An example is illustrated in (15) where the past tense particle *woon* has to be a dependent of the copula *la*. Following the UD practices, both the copular verb (e.g. *la*) and the tense particle (e.g. *woon*) have to be attached as siblings to the nonverbal predicate, as shown below.

(15) *Amari xale la woon.*
Amari child COP.3SG PAST
'Amari was a child.'



5.1.5 PRON

In WolGramBank, object and locative clitics (OLCs) are tagged as *CL* for clitics (Dione, 2013). A particular motivation for this was to distinguish these elements from subject pronouns, which are tagged as *PRON*. While subject pronouns have a predictable position in the sentence, OLCs have a quite special distribution, i.e. are special clitics according to Zwicky's definition (Zwicky, 1977).[13] First, they have a phrase structure position which is distinct from that of their non-clitic counterparts. While the latter typically follow the verb, the former usually precede it. Furthermore, OLCs have a set order amongst themselves. That is, if there is more than one clitic, they form a cluster. Considering these properties, OLCs are tagged as *CL* in WolGramBank. However, for UD compatibility reasons, both subject pronouns and object clitics are assigned the category *PRON* for pronouns. The relevant distinction is then made by using features, i.e. *Case=Nom* for subject clitics, and *Case=Acc* for object clitics. In contrast, locative clitics are assigned the *ADV* tag. Example (16) shows an instance of subject (*mu*), object (*ko*), and locative (*fa*) clitics.

(16) *Mu lekk ko fa.*
3SG.SUBJ eat 3SG.OBJ LOC
'So she/he/it eats it there.'

In addition, possessive, reflexive, relative, interrogative, demonstrative, and indefinite pronouns are also grouped under the *PRON* class. Like personal pronouns, possessive and reflexive pronouns have person and number features. Pronouns also include information about the noun class (where appropriate).

5.1.6 Adpositions

Wolof has only prepositions (no postpositions or circumpositions). The WolGramBank distinguishes between simple, partitive, and possessive prepositions. However, the UD convention does not further categorize prepositions, nor does it make a distinction between prepositions and postpositions. It rather recommends the category adposition (ADP) which is the cover term for both categories. Accordingly, in the WTB we use *ADP* without any subtype and that category actually only includes prepositions.

Table 4 shows the mapping between UD vs. WolGramBank PoS tags. It is a many-to-one (i.e. multiple WolGramBank tags mapping to one UD tag) rather than a many-to-many mapping, thus validating both annotation schemes. The WTB does not use the category *ADJ*, as the language has no adjectives.

[13] For an extensive discussion of Wolof object and locative clitics, see Zribi-Hertz and Diagne (2002).

UD PoS	Wolof Tagset	Example
ADP	PREP	*ci* 'in'
ADV	ADV	*léegi* 'now'
	CL	*fa* 'there'
AUX	AUX	*dina* 'I will'
	CL	*woon* (past tense particle)
	INFL	*a* (subj. focus marker)
CCONJ	CONJ	*ak* 'and' (nominal conjunction)
	CONJADV	*te* 'and' (clausal conjunction)
DET	DET	*bi* 'the'
	QUANT	*bépp* 'every'
INTJ	INTJ	*waaw* 'yes'
NOUN	NOUN	*kër* 'house'
	NPOSS	*këram* 'his house'
	NGEN	*këru* 'house of'
NUM	NUMBER	*fukk* 'ten'
PART	PART	*a* (infinitive particle)
PRON	PRON	*mu* (3SG subj. pron.)
	CL	*ko* (3SG obj. pron.)
PROPN	NAME	*Amari* 'Amari'
PUNCT	PUNCT	'.' period/full stop
SCONJ	COMP	*bu* 'when'
SYM	SYM	= (equal symbol)
VERB	VERB	*lekk* 'eat'
	COP	*di* 'to be'

Table 4: Mapping between the Wolof LFG and the UD PoS tagset

5.2 Morphological annotation

The UD annotation scheme defines a set of 23 morphological features across languages. These are divided into lexical vs. inflectional features. Lexical features such as *PronType* (pronoun type) and *Poss* (possessive) are attributes of lexemes or lemmas. Inflectional features are mostly features of individual word forms and are further subdivided into nominal features (e.g. *Gender*, *Case*, *Definite*) vs. verbal features (e.g. *Person*, *Number*, *Tense* and *Mood*). In contrast to the universal PoS tagset, the language specification allows treebanks to extend this set of universal features and add language-specific features when necessary.

One feature that is currently missing in the universal list of features and quite relevant for Wolof is *FocusType*. To capture the main distinction between the different focus constructions, we introduce *FocusType* as a new feature. This attribute can take three values: *subj*, *verb*, *compl* depending on the syntactic function of the constituent in focus. Another feature that needed to be updated was *NounClass*.[14] Although that feature is described in the UD guidelines, it was not used in any UD treebank so far, since UD currently does not contain any Bantu language. The description of *NounClass* indicates that the set of values of that feature is specific for a language family or group. The idea is to identify, within a language group, classes that have similar meaning across languages. However, one has to decide where the boundary of the group is.

The UD guidelines illustrate the use of the *NounClass* feature based on the system found in the Bantu language group. Following this, the feature has values that range from 1 to 20 noun classes called *Bantu1* to *Bantu20*. The class numbering system is accepted by scholars of the various Bantu languages and UD recommends the creation of similar numbering systems for the other families that have noun classes.

Because Wolof is not a Bantu language, and the Bantu classes were not extensible to Wolof, it was necessary to create a different set of classes (that could eventually be shared with some other related non-Bantu Niger-Congo languages). However, as mentioned above, one main difficulty with such an endeavour is the lack of semantic coherence in the Wolof noun class system. In most cases, and unlike in Bantu languages, there is no clear semantics, phonology or morphology that can explain the classification in Wolof.

The approach we adopted to tackle these issues was to create a set of classes for Wolof that follows a schema similar to the one proposed for Bantu languages. This means that the values of the feature had to be in a certain range (e.g. Wol1 - Wol13). It was also necessary to order the values in a way that would be comparable to the Bantu classes where possible.

[14]The NounClass feature is described in UD, since it is described in UniMorph (Sylak-Glassman, 2016).

To illustrate the numbering system in the Bantu languages, the UD guidelines listed 18 noun classes for Swahili. Some of these show a similarity with the Wolof noun classes, as illustrated in Table 5. For instance, the classes number 1 and 2 refer to singular and plural persons, respectively. It is easy to see that the Wolof equivalents of these two classes are the *k* and *ñ* class, respectively. Likewise, the classes number 7 and 8 have the typical meaning of singular and plural things, respectively. Their Wolof counterparts would be *l* and *y*, respectively. Thus, for these classes, it was not problematic to propose a comparable numbering system.

	Swahili	**Wolof**	
Class number	**Prefix**	**Affix**	**Typical Meaning**
1	*m-, mw-, mu-*	*k*	singular: persons
2	*wa-, w-*	*ñ*	plural: persons (a plural counterpart of class 1)
7	*ki-, ch-*	*l*	singular: things
8	*vi-, vy-*	*y*	plural: things (a plural counterpart of class 7)

Table 5: Noun system numbering for compatible classes between Bantu and Wolof.

However, for the remaining Wolof classes, a numbering system different from those found in Bantu was necessary. This is because the typical meaning of these Wolof classes did not match the semantics conveyed by the Bantu classes. Table 6 gives the numbering system proposed for Wolof (and eventually non-Bantu Niger-Congo languages). Also, as stated above, it is crucial to mention that the examples of typical meaning provided in this table are not meant to be reliable or systematic indicators of noun classes in Wolof. For each class, there are several words that do not follow these patterns. Also note that currently nouns are not marked with the *NounClass* feature. This is particularly motivated by the fact that nouns in Wolof (i) lack a class marker on the noun itself and (ii) may belong to several classes.

Class number	**Affix**	**Typical meaning**	**Value name**
1	*k*	singular: persons	Wol1
2	*ñ*	plural: persons	Wol2
3	*g*	singular: plants, trees	Wol3
4	*j*	singular: family members	Wol4
5	*b*	singular: fruits, default class	Wol5
6	*m*	singular: liquids	Wol6
7	*l*	singular: things	Wol7
8	*y*	plural: things	Wol8
9	*s*	singular: diminutive	Wol9
10	*w*	singular: no clear semantics	Wol10
11	*f*	locative	Wol11
12	*n*	manner	Wol12

Table 6: Noun class numbering for Wolof

As discussed in section 2.1, Wolof demonstratives encode information about deixis, including reference to the speaker and/or addressee. As with the *NounClass* feature, the *Deixis* feature is described in Unimorph (Sylak-Glassman, 2016), but not currently used by any UD treebank. So, to properly capture this information, the WTB introduced two features: *Deixis* and *DeixisRef*, which respectively represent deixis subdimensions corresponding to "Distance" and "Reference Point". The distance distinction is a three-way contrast between proximate (*Prox*), medial (*Med*), and remote (*Remt*). Reference point is used to determine the relationship of the speaker, addressee, and referent of the pronoun. The latter dimension often overlaps with distance distinctions, but is sometimes explicitly separated. In the WTB, the two primary features for reference point are speaker as reference point (ref1), and addressee as reference point (ref2). Thus, the information contained in the Wolof demonstratives given in example (3) can be modeled as follows:

- close to me, wherever you may be ... Deixis=Prox|DeixisRef=1
- far from me, wherever you may be ... Deixis=Remt|DeixisRef=1
- far from both, closer to you ... Deixis=Med|DeixisRef=2
- close to you, far from me ... Deixis=Prox|DeixisRef=2

Table 7 summarizes the morphological features used in the WTB. PoS tags that do not have additional features, e.g. coordinating conjunctions (CCONJ), subordinating conjunctions (SCONJ), interjections (INTJ), particles (PART), proper names (PROPN), punctuations (PUNCT) and symbols (SYM), are not displayed.

UD PoS	Description	Morphological Features
ADP	Adpositions	Number=Sing,Plur; NounClass=Wol1,Wol2,..,Wol13;
ADV	Adverbs	Polarity=Neg,Pos; PronType=Rel,Int
AUX	Auxiliaries	Aspect=Hab,Imp,Perf,Prog; Focus=Subj,Verb,Compl; Mood=Cnd,Imp,Ind,Opt; Number=Sing,Plur; Person=0,1,2,3; Polarity=Neg,Pos; Tense=Fut,Past,Pres; VerbForm=Fin,Inf
NOUN	Nouns	Case=Gen; Poss=Yes
DET	Determiners	Definite=Def,Ind; Deixis=Prox, Med,Remt; DeixisRef=1,2; NounClass=Wol1,Wol2,..,Wol13; Number=Sing,Plur; Poss=Yes; PronType=Art,Dem,Int,Neg,Prs,Rel,Tot
NUM	Numerals	NumType=Card,Ord
PRON	Pronouns	Definite=Def,Ind; Deixis=Prox, Med,Remt; DeixisRef=1,2; NounClass=Wol1,Wol2,..,Wol13; Number=Sing,Plur; Poss=Yes; PronType=Art,Dem,Int,Neg,Prs,Rel,Tot
VERB	Non-auxiliary verbs	Aspect=Hab; Mood=Cnd,Imp,Ind; Number=Sing,Plur; Person=0,1,2,3; Polarity=Neg,Pos; Tense=Past,Pres; VerbForm=Fin,Inf

Table 7: Morphological features in the WTB

5.3 Syntactic annotation

The WTB uses most of the UD relations, apart from *amod*, *clf*, *dep*, *goeswith*, and *reparandum*. The two first relations are not relevant for Wolof, which lacks adjectival modifier[15] and classifier. Likewise, *goeswith* and *reparandum* are not used as the WTB data do not contain dysfluencies/orthographic errors. Finally, *dep* was irrelevant as it was always possible to determine a more precise relation. Table 8 lists the frequency of UD relations used in the WTB.

UD Relation	Description	Frequency	UD Relation	Description	Frequency
acl	clausal modifier of noun	123	expl	expletive	4
acl:relcl	relative clause modifier	2336	fixed	fixed MWEs	205
advcl	adverbial clause modifier	837	flat	flat MWEs	615
advmod	adverbial modifier	1446	iobj	indirect object	298
appos	appositional modifier	298	iobj:appl	indirect applied object	7
aux	auxiliary	3301	mark	marker	1835
case	case marking	2415	nmod	nominal modifier	1821
cc	coordinating conjunction	1367	nsubj	nominal subject	4395
ccomp	clausal complement	733	nummod	numeric modifier	377
compound	coumpound	220	obj	object	3318
compound:prt	phrasal verb particle	68	obj:appl	applied object	76
compound:svc	serial compound verb	75	obj:caus	causative object	118
conj	conjunction	1877	obl	oblique nominal	2138
cop	copula	626	obl:appl	applied oblique	79
csubj	clausal subject	50	orphan	orphan	13
det	determiner	3138	parataxis	parataxis	412
discourse	discourse elements	47	punct	punctuation	5319
dislocated	dislocated elements	548	xcomp	open clausal complement	928

Table 8: Universal dependency relations in WTB

6 Conclusion

This paper has presented the process of creating a Universal Dependency treebank for Wolof, the first UD treebank from the North Atlantic languages. Wolof is also the second Atlantic-Congo language (after Yoruba) that has a UD treebank. Adopting UD to existing conventions for annotating Wolof required several decisions to be made. We have discussed issues related to tokenization pointing out the challenge of clitic segmentation. We indicated that Wolof orthographic words may carry morphological information as well as other function elements of syntactic relations. The discussion has also shown that there are a number of challenges in adapting the UD scheme for Wolof. In particular we advocate the introduction of missing features for focus marking and deixis information, and the redefinition of the existing noun class feature for non-Bantu languages. In future, we plan to address the issue of automatic conversion of WolGramBank.

[15]The *amod* relation is only used to annotate foreign material (e.g. French texts) that is contained in the WTB.

Acknowledgements

I would like to thank the UD community, in particular Dan Zeman for many fruitful discussions. I also want to thank the anonymous reviewers for valuable comments and suggestions.

References

Anne Abeillé, editor. 2003. *Treebanks: Building and Using Parsed Corpora*. Kluwer.

Mohammed A Attia. 2007. Arabic Tokenization System. In *Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources*, pages 65–72. Association for Computational Linguistics.

Eric D. Church. 1981. *Le système verbal du wolof*. Faculté des Lettres et Sciences Humaines (FLSH), Université de Dakar.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.

Cheikh M. Bamba Dione. 2012a. A Morphological Analyzer For Wolof Using Finite-State Techniques. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. ELRA.

Cheikh M. Bamba Dione. 2012b. An LFG Approach to Wolof Cleft Constructions. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG '12 Conference*, Stanford, CA. CSLI Publications.

Cheikh M. Bamba Dione. 2013. Handling Wolof Clitics in LFG. In Christine Meklenborg Salvesen and Hans Petter Helland, editors, *Challenging Clitics*, Amsterdam. John Benjamins Publishing Company.

Cheikh M Bamba Dione. 2014. LFG parse disambiguation for Wolof. *Journal of Language Modelling*, 2(1):105–165.

Cheikh M. Bamba Dione. 2017. Finite-state tokenization for a deep wolof lfg grammar. *Bergen Language and Linguistics Studies*, 8(1).

Kim Gerdes. 2013. Collaborative dependency annotation. In *Proceedings of the second international conference on dependency linguistics (DepLing 2013)*, pages 88–97.

Joseph H Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.

Alain Kihm. 2000. Wolof Genitive Constructions and the Construct State. In J. Lowenstamm & U. Shlonsky Lecarme, J., editor, *Research in Afro-Asiatic grammar: papers from the third conference on Afroasiatic languages*, pages 150–181. Amsterdam & Philadelphia : John Benjamins Publishing Co.

Christopher D Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Fiona McLaughlin. 1997. Noun classification in Wolof: When affixes are not renewed. *Studies in African Linguistics*, 26(1).

Fiona McLaughlin. 2004. Is there an adjective class in Wolof? In R.M.W. Dixon and Alexandra Y. Aikhenvald, editors, *Adjective classes. A crosslinguistic typology.*, pages 242–262. Oxford University Press.

Paul Meurer. 2017. From LFG structures to dependency relations. *Bergen Language and Linguistics Studies*, 8(1).

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Adam Przepiórkowski and Agnieszka Patejuk. 2019. From lexical functional grammar to enhanced universal dependencies. *Language Resources and Evaluation*, Feb.

Stéphane Robert. 1991. Approche énonciative du système verbal: le cas du wolof. *Editions du CNRS*.

Stéphane Robert. 2000. Le verbe wolof ou la grammaticalisation du focus. Louvain: Peeters, Coll. Afrique et Langage, 229-267. Version non corrigée.

Stéphane Robert. 2010. Clause chaining and conjugations in wolof. *Clause Linking and Clause Hierarchy: Syntax and Pragmatics*, 121:469–498.

Stéphane Robert. 2016. Content question words and noun class markers in wolof: reconstructing a puzzle. *Frankfurt African Studies Bulletin*, 23:123–146.

Antoine de Saint-Exupéry. 1971. Le petit prince. 1943. *Paris: Harvest*.

David J. Sapir. 1971. West Atlantic: an inventory of the languages, their noun class systems and consonant alternation. *Current Trends in Linguistics*, 7(1):43–112.

Binyam Ephrem Seyoum, Yusuke Miyao, and Baye Yimam Mekonnen. 2018. Universal dependencies for amharic. In *LREC*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Sebastian Sulger, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Agnieszka Patejuk, Özlem Çetinoğlu, I Wayan Arka, and Meladel Mistica. 2013. ParGramBank: The ParGram Parallel Treebank. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 759–767, Sofia, Bulgaria.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema). Technical report, Technical report, Department of Computer Science, Johns Hopkins University.

Khady Tamba, Harold Torrence, and Malte Zimmermann. 2012. Wolof quantifiers. In *Handbook of Quantifiers in Natural Language*, pages 891–939. Springer.

Lucien Tesnière. 1959. Eléments de syntaxe structurale. *Klincksieck, Paris*.

Francis M Tyers, Mariya Sheyanova, and Jonathan North Washington. 2018. UD Annotatrix: An annotation tool for universal dependencies. In *Proceedings of the 16th Conference on Treebanks and Linguistic Theories*.

Anne Zribi-Hertz and Lamine Diagne. 2002. Clitic placement after syntax: evidence from Wolof person and locative markers. *Natural Language & Linguistic Theory*, 20(4):823–884.

Arnold Zwicky. 1977. On Clitics. *Indiana University Linguistics Club*.