Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5128–5136
Marseille, 11–16 May 2020

Implementation and Evaluation of an LFG-based Parser for Wolof

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

Abstract

This paper reports on a parsing system for Wolof based on the LFG formalism. The parser covers core constructions of Wolof, including noun classes, cleft, copula, causative and applicative sentences. It also deals with several types of coordination, including same constituent coordination, asymmetric and asyndetic coordination. The system uses a cascade of finite-state transducers for word tokenization and morphological analysis as well as various lexicons. In addition, robust parsing techniques, including fragmenting and skimming, are used to optimize grammar coverage. Parsing coverage is evaluated by running test-suites of naturally occurring Wolof sentences through the parser. The evaluation of parsing coverage reveals that 72.72% of the test sentences receive full parses; 27.27% receive partial parses. To measure accuracy, the parsed sentences are disambiguated manually using an incremental parsebanking approach based on discriminants. The evaluation of parsing quality reveals that the parser achieves 67.2% recall, 92.8% precision and an f-score of 77.9%.

Keywords: Parser, LFG, low-resource language, treebank, Wolof.

1. Introduction

Deep grammars that follow an established linguistic theory such as Lexical Functional Grammar (LFG) (Bresnan, 2001) provide detailed syntactic analysis that is essential for the further development of NLP applications. This paper reports on the development and evaluation of the first LFG parsing system for Wolof, a low-resource Niger-Congo language mostly spoken in Senegal. The system is developed as part of the Parallel Grammar (ParGram) project (Butt et al., 2002) and is based on the Xerox Linguistic Environment (XLE) (Crouch et al., 2019). The goal of this research work is to provide a practical parsing system with broad coverage and deep analysis of naturally occurring Wolof data. The system described in this paper is the first parser reported for Wolof.
LFG assumes two core levels of syntactic analysis: a c(onstituent)-structure which characterizes the phrase structure configurations as a phrase structure tree, and a f(unctional)-structure which encodes grammatical relations (e.g. subject, object) and features (e.g. person, number).
For instance, when coupled with XLE, the Wolof grammar assigns to example (1) the c- and f-structure in Figure 1.[1]

(1) *Janq y-i bind na-ñu téeré b-i.*
girl NC-P write FIN-3PL book NC-P
"The girls wrote the book."

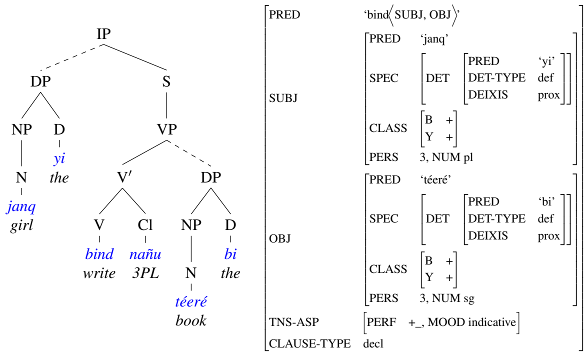

Figure 1: C- and f-structure of sentence (1)

The c-structure organization proposed for Wolof is briefly discussed in section 2.2.. The f-structure in Figure 1 indicates that the main predicate of (1) is *bind* 'write' and has a subject (SUBJ) and an object (OBJ). SUBJ has a semantic predicate (*janq*) and is analyzed as a noun that belongs to the *b* and *y* noun classes (see section 2.1.). The SPEC feature in the SUBJ f-structure is introduced by the definite (def) determiner *yi*, which has a semantic predicate (*yi*) and encodes deixis information (*proximal*). OBJ shows a similar f-structure. The sentence is an indicative declarative clause expressed in the perfective aspect through the combination of the *na* morpheme and the lexical verb.
This paper is organized as follows. Section 2. highlights some key issues in Wolof and describes how these are addressed in the grammar. Section 3. presents the general architecture of the Wolof parsing system. Section 4. describes the data used for parser evaluation. Section 5. reports on the results of the experimental evaluation. Section 6. discusses issues related to the parsing coverage and quality. Finally, section 7. concludes the discussion.

2. Wolof Morphosyntax

2.1. Noun classes

Wolof has a noun class (NC) system with 8 singular and 2 plural noun classes (McLaughlin, 1997). Class membership is typically expressed by a class index on nominal dependents such as determiners and relative pronouns rather than on the noun itself. The indexes for singular noun classes are: *b*, *g*, *j*, *k*, *l*, *m*, *s*, *w*, and for plural noun classes: *y*, and *ñ*. The *k* and *ñ* classes are human classes, while *l* and *y* are typically non-human classes. However, the Wolof NC system generally lacks semantic coherence (McLaughlin, 1997).
Determiners agree in NC with their head noun. The NC is indicated by a word-initial consonant in the definite article,

[1] Abbreviations in the glosses: COP: copula; D: distal; FIN: finite; IPFV: imperfective; NC: noun class; NSFOC: non-subject focus; P: proximal; PL: plural; SG: singular; SFOC: subject focus; SUBJ: subject; VFOC: verb focus; 1, 2, 3: 1st, 2nd, 3rd person.

which encodes deixis regarding the noun reference. The suffixes *-i* and *-a* in (2a) and (2b) signal that the article is definite proximal (P) and distal (D), respectively.

(2) a. *janq b-i*
girl NC-P
'the girl (proximal)'

b. *janq b-a*
girl NC-D
'the girl (distal)'

A key issue with the treatment of NCs in the Wolof grammar is lexical ambiguity (Dione, 2014). Due to homonymy or polysemy, a noun may belong to many classes. For instance, the same noun form *ndaw* can occur with five classes: *g* (e.g. *ndaw gi* "the youth"), *l* (e.g. *ndaw li* "the messenger"), *s* (e.g. *ndaw si* "the young woman"), *ñ* (e.g. *ndaw ñi* "the young people/women") and *y* (e.g. *ndaw yi* "the messengers").

The lexical ambiguity highlights the fact that the Wolof NCs illustrate as case of feature indeterminacy, as has been observed for other languages (Dalrymple et al., 2009). For instance, the German plural noun form *Papageien* 'parrots' shows no CASE distinction and can meet different CASE requirements (i.e. nominative, accusative, dative, genitive). Likewise, noun forms in Wolof show no overt noun class distinction, and thus can meet different class requirements. Accordingly, the Wolof NCs are treated similar to the representation of CASE in German. Thus, a noun like *ndaw* has the NC attribute in (3), whose value specifies each noun class by means of a separate boolean-valued attribute: *G*, *L*, *S*, *Ñ*, and *Y*. Nouns and their modifiers specify negative values or do not specify any value for the noun classes they do not express, and specify or are compatible with positive values for the classes they do express (Dione, 2014).

(3)
$$\left[\begin{array}{ll}\text{CLASS} & \left[\begin{array}{ll}\text{G} & + \\ \text{L} & + \\ \text{S} & + \\ \text{Ñ} & + \\ \text{Y} & +\end{array}\right]\end{array}\right]$$

2.2. Verbal syntax and clausal organization

In Wolof, verbal inflection is typically not marked on the verb itself, but rather carried out by special markers, which express grammatical specifications of the verb, including person, number, tense, aspect, mood and focus (Robert, 2000). The inflectional markers can be preposed, postposed, or suffixed to the lexical stem, resulting in several complex clause types. For instance, the imperfective (IPFV) form of (1) can be expressed using the *di* auxiliary (4).

(4) *Janq y-i di-na-ñu bind téeré b-i.*
girl NC-P IPFV-FIN-3PL write book NC-P
"The girls will write the book."

A crucial property of the inflectional markers is their ability to express information structure (Robert, 2000; Dione, 2012b). In fact, Wolof has morphosyntactic means to mark focus on the subject, verb, or non-subject constituent (i.e. any constituent which is neither subject nor verb), as shown in (5a), (5b) and (5c), respectively. Morphologically, the origins of the subject, verb and non-subject focus markers are *-a*, *da-* and *la-*, respectively.

(5) a. *Janq y-i ñu-a (>ñoo) bind téeré b-i.*
girl NC-P 3PL-SFOC write book NC-P
"It's the girls who wrote the book."

b. *Janq y-i da-ñu bind téeré b-i.*
girl NC-P VFOC-3PL write book NC-P
"What the girls did is write the book."

c. *Téeré b-i la janq y-i bind.*
book NC-P NSFOC.3 girl NC-P write
"It's the book that the girls wrote."

The focus marker takes a different form depending on the focus type, the person and number of the subject. Moreover, the marker precedes the focused constituent in verb focus, but follows it in subject and non-subject focus clauses.

The clausal syntax of Wolof suggests that the language exhibits a mixture of an endocentric and exocentric organization. Grammatical functions are often encoded through phrase structure position, but in some clauses, they must be localized by means of morphology. This is evidenced by the typology of the non-subject focused clauses (5c).

Thus, the c-structure organization proposed for Wolof (Dione, 2013a) identifies a sentence with *IP*, which is the projection of I^o (for inflection). This captures the generalization that the finite auxiliary, e.g. *dinañu* in (4), and other inflectional elements (e.g. the focus markers) occupy a unique position in the sentence. *IP* may consist of a nominal specifier, e.g. a *DP* (determiner phrase) as in Figure 1, and the exocentric category *S*. The specifier of *IP* may link to SUBJ, or to different grammatical functions (e.g. OBJ, OBJ-TH, COMP, XCOMP, ADJUNCT) in non-subject focus clauses. The exocentric category *S* has no fixed head and is assumed for non-configurational structures (Bresnan, 2001) like (5c). A *DP* typically consists of a noun phrase (*NP*/*N*) and a determiner (*D*).

Figure 2 shows the c-structure and simplified f-structure associated with (5c). The subject appears under *S*, while the specifier of *IP* simultaneously bears the OBJ and FOCUS functions, as indicated by the shared index 7 in the f-structure.

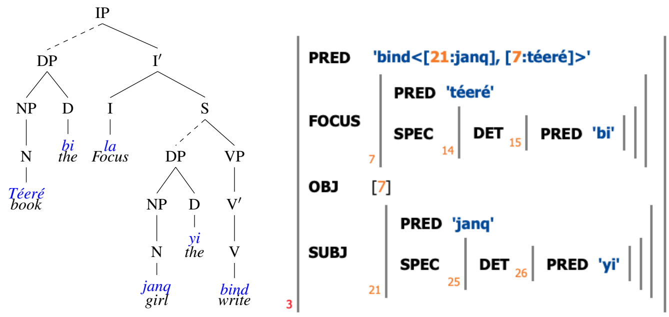

Figure 2: C- and f-structure of sentence (5c)

Inflectional markers that are always immediately postposed to the verb like *nañu* in (1) are affixes. Therefore, they are not treated as instances of I^o, but rather as a special category (Cl) that is a daughter node of V'. Other nodes that may appear under V' as sister of these affixes are object and locative clitics (not discussed here), which are different form affixes in the sense that they are syntactic clitics.

2.3. Copula Constructions

In Wolof, copula constructions are somewhat related to the focus sentences in that both may instantiate the same form. For instance, verb focus sentences (5b) and verb copulas (6a) use the same form (*da*); and similarly for non-subject focus (5c) and non-subject copula (6b) regarding the *la* form.

(6) a. *Janq y-i da-ñu rafet.*
girl NC-P COP-3PL be.beautiful
"The girls are beautiful / BEAUTIFUL"

b. *Janq y-i bindkat la-ñu.*
girl NC-P writer COP-3PL
'The girls are writers.'

The major challenge to modeling these constructions is the lack of a uniform analysis of copula in LFG. Instead, three different approaches can be identified (Dalrymple et al., 2004; Attia, 2008): a single-tier analysis, an open-complement double-tier analysis and a closed complement double-tier analysis.

In the single-tier approach, the copula predicate (i.e. the adjective "red" in a sentence like "the car is red") functions as the sentential head and selects for a subject. This approach is recommended for cases like Japanese predicative adjectives in which the copula is optional. Unlike Japanese, Wolof does not have the adjective category (McLaughlin, 2004). It rather uses stative verbs to express the 'adjectival' concept, and these behave similar to Japanese adjectives in that: (i) they license their own subject and (ii) they do not require the copula (7). When the copula is present (6a), this may result in focalization (see the English translation of (6a)). The single-tier analysis is adopted for Wolof stative verbs, as the simplified f-structure of (6a) in (8) shows.

(7) *Janq y-i rafet na-ñu.*
girl NC-P be.beautifful FIN-3PL
"The girls are beautiful."

(8)
$$\left[\begin{array}{ll} \text{PRED} & \text{`rafet}\left\langle\text{SUBJ}\right\rangle\text{'} \\ \text{SUBJ} & \left[\begin{array}{ll}\text{PRED} & \text{`janq'}\end{array}\right] \\ \text{FOCUS} & \left[\text{rafet}\right] \end{array}\right]$$

However, the assumption that the copula predicate selects for a subject is problematic for e.g. NPs or PPs which don't. In contrast, the first variant of the double-tier analysis follows the earliest treatments of copulas in LFG (Bresnan, 1982). In this approach, the copula predicate functions as an open complement (XCOMP-PRED) whose subject raises to the matrix clause as a non-thematic subject of the copula *be*. For instance, the sentence "the girls are writers" is analyzed in the English ParGram grammar as shown in Figure 3.

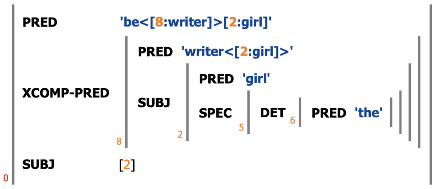

Figure 3: English copula example

This approach, however, faces the same issue as the previous one in that it assumes that the copula predicate is open, and therefore must subcategorize for a subject. Again, this is problematic for phrasal constituents like NPs and PPs which do not have an overt subject. Moreover, the open complement analysis results in a clash of PRED values if the post-copular complement has a subject.

In the second variant of the double-tier analysis, the copula predicate functions as a closed complement of the copula (PREDLINK). This eliminates the need for a control equation between the subject and the copula predicate, solving the issues mentioned above. Furthermore, the PREDLINK analysis is not affected by the constituent type of the copula complement, i.e. it can handle any constituent types with different semantic roles. Accordingly, the Wolof grammar uses this approach for copula constructions like (6b), in which the copula predicate is typically nonverbal. The simplified f-structure associated with (6b) is given in Figure 4. As the shared index 2 shows, the copula predicate may function as FOCUS. Likewise, in such constructions, the subject position is typically associated with TOPIC.

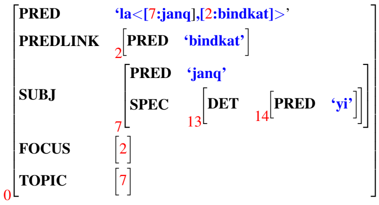

Figure 4: Double-tier analysis (PREDLINK) of (6b)

In ParGram, the lack of a uniform analysis of copula is seen as a way to account for the typological differences found with copulas across languages (Sulger et al., 2013).

2.4. Causatives

Wolof morphotactics allows the derivation of causative forms for verbs by means of different suffixes (Comrie, 1985; Dione, 2013b). Example (9) shows causative forms of the sentence in (1). In (9a), the suffix *-loo* signals the addition of a new subject argument, which semantically is the causer of the action. Likewise, the morpheme *-lu* in (9b) introduces the causer as SUBJ, but unlike *-loo*, reduces the object position by removing the former subject (the causee).

(9) a. *Janq y-i bind-loo na-ñu Awa téeré b-i.*
girl NC-P write-CAUS FIN-3PL Awa book NC-P
"The girls made Awa write the book."

b. *Janq y-i bind-lu na-ñu téeré b-i.*
girl NC-P write-CAUS FIN-3PL book NC-P
"The girls let (someone) write the book."

Causative can also be derived by means of the suffix *-al* (10b). This suffix only attaches to unaccusative verbs (i.e. verbs with a patient subject) to express transitive causative.

(10) a. *Mburu m-i tooy na.*
bread NC-P be.wet 3SG
"The bread is wet."

b. *Awa tooy-al na mburu m-i.*
Awa be.wet-CAUS 3SG bread NC-P
"Awa made the bread wet."

The simplified f-structures related to the causative sentences (9a-9b) are shown in Figures 5 and 6, respectively.

$$\left[\begin{array}{ll}
\text{PRED} & \text{`caus}\left\langle\text{janq, `bind}\left\langle\text{Awa, téeré}\right\rangle\text{'}\right\rangle\text{'} \\
\text{SUBJ} & \left[\begin{array}{ll}\text{PRED} & \text{`janq'} \\ \text{SPEC} & \left[\text{DET} \left[\text{PRED} \quad \text{`yi'}\right]\right]\end{array}\right] \\
\text{OBJ-TH} & \left[\begin{array}{ll}\text{PRED} & \text{`téeré'} \\ \text{SPEC} & \left[\text{DET} \left[\text{PRED} \quad \text{`bi'}\right]\right]\end{array}\right] \\
\text{OBJ} & \left[\text{PRED} \quad \text{`Awa'}\right] \\
\text{CAUSATIVE} & +
\end{array}\right]$$

Figure 5: F-structure of (9a)

$$\left[\begin{array}{ll}
\text{PRED} & \text{`caus}\left\langle\text{janq, `bind}\left\langle\text{NULL, téeré}\right\rangle\text{'}\right\rangle\text{'} \\
\text{SUBJ} & \left[\begin{array}{ll}\text{PRED} & \text{`janq'} \\ \text{SPEC} & \left[\text{DET} \left[\text{PRED} \quad \text{`yi'}\right]\right]\end{array}\right] \\
\text{OBJ} & \left[\begin{array}{ll}\text{PRED} & \text{`téeré'} \\ \text{SPEC} & \left[\text{DET} \left[\text{PRED} \quad \text{`bi'}\right]\right]\end{array}\right] \\
\text{CAUSATIVE} & +
\end{array}\right]$$

Figure 6: F-structure of (9b)

The Wolof causatives are treated as complex predicates (Butt, 1995). Their argument structure consists of the matrix predicate *caus* (for causative) which has the causer as SUBJ and the original verb (e.g. *bind*) as embedded predicate. The difference in terms of valency change between the suffixes *-loo* and *-lu* can be observed in the f-structures. If the causative morpheme is *-loo*, the subject and object of the original verb become OBJ and OBJ-TH (i.e. secondary object), respectively (Fig. 5). In contrast, if the causative morpheme is *-lu*, the original verb is assumed to have a null SUBJ (reflecting the removal of the former subject); the former direct OBJ remains unchanged (Fig. 6).

2.5. Applicatives

Likewise, Wolof morphology allows the production of applicative suffixes (e.g. *-al* and *-e*) to code different semantic roles. For instance, compared to (1), the suffix *-al* in (11) signals that a new object argument (*Awa*) with the semantic role beneficiary has been added.

(11) *Janq y-i bind-al na-ñu Awa téeré b-i.*
girl NC-P write-APPL FIN-3PL Awa book NC-P
"The girls wrote the book for Awa."

Note the applicative-causative polysemy in Wolof: the same suffix may be used to derive both causative and applicative. This is true for both suffixes *-al* and *-e*. For instance, in (10b) the suffix *-al* does not signal applicative morphology, but rather causativization of the verb *tooy* "to be wet".
The applicative suffix *-e* licenses objects with a semantic role of instrumental (12), locative or manner. As a causative suffix, *-e* is lexicalized and limited to e.g. unergative verbs, which have an agent subject like *génn* (13).

(12) *Awa togg-e na jën w-i diw.*
Awa cook-APPL FIN.3SG fish NC-P oil
"Awa cooked the fish with oil."

(13) *Awa génn-e na jën w-i.*
Awa go.out-CAUS FIN.3SG fish NC-P
"Awa let/made the fish go out."

This parallelism between causative and applicative in Wolof suggested a unified approach for these types of constructions. Thus, like the causative constructions discussed above, applicative sentences are treated as complex predicates. For instance, the sentence (11) is analyzed as having a matrix and an embedded predicate, as its simplified f-structure in Figure 7 shows.

$$\left[\begin{array}{ll}
\text{PRED} & \text{`appl}\left\langle\text{`bind}\left\langle\text{janq , téeré}\right\rangle\text{', Awa}\right\rangle\text{'} \\
\text{SUBJ} & \left[\begin{array}{ll}\text{PRED} & \text{`janq'} \\ \text{SPEC} & \left[\text{DET} \left[\text{PRED} \quad \text{`bi'}\right]\right]\end{array}\right] \\
\text{OBJ-TH} & \left[\begin{array}{ll}\text{PRED} & \text{`téeré'} \\ \text{SPEC} & \left[\text{DET} \left[\text{PRED} \quad \text{`bi'}\right]\right]\end{array}\right] \\
\text{OBJappl} & \left[\text{PRED} \quad \text{`Awa'}\right] \\
\text{APPLICATIVE} & +
\end{array}\right]$$

Figure 7: F-structure of (11)

Unlike with causatives, the embedded predicate in applicative derivation occupies the first position in the matrix predicate, while the applied object (*Awa*) takes the second position. The clause is analyzed as having a displaced theme (i.e. *téeré* 'book). According to Bresnan and Moshi (1990), a theme is ranked low in the thematic hierarchy. Therefore, the displaced theme is mapped to the OBJ-TH function, as we assume that the other prominent arguments are realized as SUBJ and direct object (OBJ), respectively. In the Wolof grammar, the newly introduced OBJ is referred to as OBJappl to make it clear that it is an applied object.

2.6. Coordination

The Wolof grammar covers several types of coordination, including same constituent coordination (SCC), asymmetric and asyndetic coordination. SCC is exemplified in (14) (here a coordination of two sentences). Using the standard LFG approach to coordination (Kaplan and Maxwell, 1988), the c-structure rule (15) allows to model SCC straightforwardly.[2] In the associated f-structure (not shown here), the coordination is represented as a set-valued f-structure where each of the conjuncts represents an element within the set, by the functional annotations $\uparrow \in \downarrow$. *XP* is a variable that can stand for any phrasal constituent.

(14) *Móodu ñëw walla Awa dem.*
Móodu come or Awa go
"Either Móodu comes or Awa leaves."

(15) $\begin{array}{lllll} \text{XP} & \rightarrow & \text{XP} & \text{CONJ} & \text{XP} \\ & & \uparrow\in\downarrow & \uparrow=\downarrow & \uparrow\in\downarrow \end{array}$

To account for asymmetric coordination like a coordination of ADVPs and PPs (16) that function as modifiers, we define special rules like (17).

(16) *Fii ak ca dëkk b-a*
here and in county NC-D
"Here and in the county"

(17) $\begin{array}{lllll} \text{ADVP} & \rightarrow & \text{ADVP} & \text{CONJ} & \text{PP} \\ & & \uparrow\in\downarrow & \uparrow=\downarrow & \uparrow\in\downarrow \end{array}$

[2] For simplicity of presentation, we only present simplified versions of the grammar rules.

Likewise, coordination of nominals is handled by a special rule (18), which includes scope relation stated as ↓ ∈<h>s ↑ . This annotation encodes the head-precedence order relation between the f-structure set elements. It is particularly relevant for handling structures like (19) where the conjuncts have different person features (first and third person, respectively), but the person feature of the set as a whole needs to be resolved according to the feature of the first conjunct.

(18) NOMCoord → { NOM: ↓ ∈<h>s ↑ ;
CONJ: @PERS-FEAT;
NOM: ↓ ∈<h>s ↑
}.

(19) *Man ak Awa bind na-nu téeré b-i.*
1SG CONJ Awa write FIN-1PL book NC-P
"Faatu and I wrote the book."

The invocation @PERS-FEAT in (18) refers to the template (20), which provides the correct person features. Templates are a shorthand used to state generalizations that need to apply to large sections of the grammar or lexicon (Butt et al., 1999). The template in (20) makes use of *if-else* logical operators to force the person to be first whenever one of the conjuncts is first person. Otherwise, if a conjunct is second person and the set is not already first person, it becomes second person. If none of these cases match, then the set must be third person.

(20) PERS-FEAT = @(IFELSE (↑ ∈ PERS)$=_c$ 1
(↑ PERS)=1
@(IFELSE (↑ ∈ PERS)$=_c$ 2
(↑ PERS)=2
(↑ PERS)=3
)
).

The Wolof data contains many instances of asyndetic coordination with subject gap, as illustrated in (21). The two conjuncts are coordinated without an explicit conjunction.

(21) *Ca dëkk b-a la Awa dem bind téeré b-i.*
in county NC-D NSFOC.3 Awa go write book NC-P
"It's to the county that Awa went and wrote the book."

Crucially, the standard approach to coordination does not allow to directly model asyndetic coordination. First, the subject in constructions like (21), i.e. *Awa*, is realized within the first conjunct. This means that it is not distributed to the second conjunct (i.e. it is a missing SUBJ function), violating *Completeness* (Bresnan, 2001, p. 63). Moreover, there may be a distribution of arguments of the first conjunct which are not subcategorized for by the second conjunct, violating *Coherence* (ibid).

The approach to asyndetic coordination with subject gap adopted for Wolof follows a symmetric analysis with *asymmetric grammaticalised discourse function (GDF) projection*, as proposed for German subject-gap constructions (Frank, 2002). In Wolof, the GDF functions are defined as the class of functions that occupy the specifier position of IP and S (e.g. SUBJ, FOCUS). For instance, (22) defines S coordination in c-structure, with symmetric projection of the conjunct's f-structures in terms of the classical ↓ ∈ ↑ annotations. Here SUBJ is the instantiated GDF. The annotation (↑ SUBJ)=(↓ SUBJ) defines the first conjunction's subject as the subject of the coordination as a whole.

(22) S → { S: ↓ ∈ ↑ (↑ SUBJ) = (↓ SUBJ);
e: (↑ COORD-FORM)=null (↑ COORD) =+;
S: ↓ ∈ ↑
}.

3. The Wolof Parsing System

At the current state, the Wolof grammar has 250 XLE rules (with regular expression-based right-hand sides) which compile into an automaton with 2737 states and 39189 arcs. Besides the grammar rules, the main components of the parser include finite-state transducers (FST) (Kaplan et al., 2004) for tokenization and morphological analysis, and LFG lexicons, as briefly discussed in the following sections.

3.1. FST Tokenizer

During preprocessing, the parser uses a cascade of FSTs. The first one acts as a tokenizer and a normalizer (Dione, 2017). It splits the input stream into a unique sequence of tokens separated by whitespaces (e.g. space, line break) or by punctuation characters. For sentences that only contain words that are clearly separated by whitespaces, tokenization is quite straightforward. However, in many other cases, tokenization faced non-trivial issues that require language-specific information. These include word contraction observed in cliticization and multiword expressions (MWE). For instance, the underlined word *yeek* in (23) is a contracted form (*yi*+*ak*) of the determiner *yi* and the coordinating conjunction *ak* "and". The surface form *yeek* is the result of vowel coalescence. Furthermore, the double-underlined sequence of words in (23) is a MWE that translates into English as "the university" (lit.: "the school which is high").

(23) *Janq yeek Awa gis nañu Daara ju Kawe ji*
girl NC.P.CONJ Awa see 3PL University
'The girls and Awa saw the University.'

Figure 8 shows how the contracted form, i.e. "yeek" in the nominal coordination (NOMCoord), and the MWE in (23) are parsed correctly by the Wolof grammar. In the c-structure, the former is analyzed as individual tokens (*yi* and *ak*), while the latter is treated as a single unit.

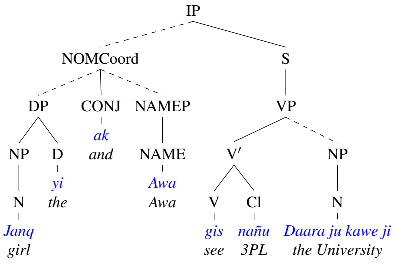

Figure 8: Phrase structure of (23)

3.2. The Morphological Analyzer

The second preprocessing component is the Wolof Morphological Analyzer (WoMA) (Dione, 2012a). Based on the Xerox finite-state tool, *fst* (Beesley and Karttunen, 2003), WoMA uses two-level representations to handle the input in both directions: analysis and generation. All inflected forms of the same word (i.e. surface forms) are mapped to the same canonical dictionary form (lemma) followed by a set of tags encoding the morphological features. The morphological analyses for the non-derived verb *bind* (24a) and for its causative (24b-24c) and applicative (24d) forms is given in (24).

(24) a. bind ⇔ bind+Verb+Base+Main+Active
b. bindloo ⇔ bind+Caus+LOO+Verb+Base+Main+Active
c. bindlu ⇔ bind+Caus+LU+Verb+Base+Main+Active
d. bindal ⇔ bind+Appl+AL+Verb+Base+Trans+Active

Example (24a) suggests that the surface form *bind* can be analysed as the base form of a main active verb. The tag +*Active* is used in contrast to +*Stative* for stative verbs, as the distinction is important for the treatment of copula and tense-aspect in Wolof. The entries (24b-24c) respectively indicate that the corresponding item is a causative form of bind derived by means of *loo* and *lu*. Likewise, (24d) conveys information about the applicative derivation.

(25) a. janq ⇔ janq+Noun+Comm+b+y
b. yi ⇔ yi+Det+Def+y+3+Pl+Prox
⇔ yi+Pron+ProSyn+Rel+3+y+Pl+Prox

Example (25a) states that the surface form *janq* "girl" can be analyzed as a stem *janq*. The tags +*Noun* and +*Comm* respectively indicate the part-of-speech (POS) and the basic syntactic type of the noun as a common noun. The noun agrees with its dependents in the *b* and *y* classes. As discussed in section 2.1., all class indexes compatible with this form should be contained in the morphological output.

The entry in (25b) illustrates cases where a single form corresponds to more than one analysis.[3] The first analysis handles the form *yi* as a definite (+Def) determiner (+Det) that is compatible with the *y* noun class (+y). In addition, this form is inflected for person (+3) and plural (+Pl). It also encodes deixis information (+Prox for proximal). The second analysis is similar to the first one, except that the form is treated as a relative (+Rel) pronoun (Pron) rather than a determiner. The tag +ProSyn indicates that the syntactic type is a pronoun (in contrast to common nouns).

3.3. Interfacing Morphology with Syntax

WoMA is interfaced with XLE by means of sublexical rules (see Kaplan and Newman (1997)), which parse the output of the morphology. For instance, the output in (25a) is treated as a phrase, and as such, each of its items is encoded in the lexicon, as one would do with any lexical entry (26). Due to lack of space, only the analysis of some items is displayed.

(26)	janq	N-S	XLE	(↑ PRED)='janq'.
	+Noun	N-TAG	XLE	.
	+b	PRON-SFX	XLE	(↑ CLASS B)=+;
		NUM-SFX	XLE	(↑ CLASS B)=+;
		D-SFX	XLE	(↑ SPEC DET CLASS B)=+;
		Q-SFX	XLE	(↑ SPEC QUANT CLASS B)=+;
		V-SFX	XLE	(↑ ARG CLASS B)=+.

The structure of each entry in (26) consists of four columns. The first and the second columns respectively indicate the base form and the category or POS tag associated with the item (e.g. N-S for noun stem). A given item may belong to several categories. For instance, +*b* can be a suffix (SFX) that attaches to e.g. pronouns (*PRON-SFX*), numbers (*NUM-SFX*), determiners (D-SFX), quantifiers (Q-SFX) and verbs (V-SFX). The XLE tag in the third column indicates that morphological information comes from the Wolof FST.

The fourth column shows a list of possible attributes and values (may be empty). For instance, the +*b* tag specifies the noun class of the corresponding f-structure. This information may be encoded at the top f-structure (e.g. for pronouns), or deeper (e.g. under the SPEC DET f-structure for determiners). In few cases, the noun class of a verb argument (e.g. OBJ, OBJ-TH, PREDLINK) may be marked on the verb itself. In such cases, this feature is provided through a functional annotation that refers to the f-structure of that argument (ARG). To simplify the reading of the lexicon, ca. 75 XLE templates for lexical entries have been implemented.

3.4. The LFG Lexicons

For modularity and transparency reasons, the parsing system uses three lexicon files: (i) a lexicon file of semi-automatically generated verb and noun stems, (ii) a lexicon file containing core entries that belong to a closed class; and (iii) a lexicon handling sublexical tags used by WoMA.

The first lexicon serves as a record for information about verb subcategorization. It contains ca. 2000 verb stems and 2836 subcategorization frame-verb stem entries. It also carries noun entries that typically consist of the noun stem or lemma and optionally the gloss. An important number of nouns and all other information (e.g. related to noun classes) come direct from the morphology or are eventually guessed. The second lexicon contains closed class items such as stems for determiners, pronouns, prepositions, etc. The third lexicon deals with the sublexical tags that are produced by the Wolof morphological analyzer, as illustrated with the +*b* tag in (26). This lexicon also includes complex predicates entries such as morphological applicative and causative.

Unknown lexical entries (those words recognized by WoMA but not found in the lexicons and those not recognized by WoMA at all) are guessed using different strategies. For instance, many lexical items have entirely predictable subcategorization frames. For these, the knowledge about the part-of-speech and some inflectional information may be sufficient for determining the lexical entry (Kaplan et al., 2004). The guessing mechanisms were helpful for recognizing nouns, adverbs and numbers.

4. Treebank and Evaluation Data

The development and testing of the grammar is based on a corpus of natural Wolof texts (short stories (Cissé, 1994; Garros, 1997) and a semi-autobiographical novel (Ba, 2007)). The advantage of using this corpus is twofold: (i) it contains heterogenous texts from different genres, sentence length and size; and (ii) it consists of real-life data with a moderate level of complexity and variation. This has helped to promote the expansion of the grammar rules and lexicons.

[3] All these forms are passed to the grammar for parsing in order to avoid early pruning of potentially correct analyses.

The basic development set consists of 380 randomly selected sentences from the short stories (Cissé, 1994; Garros, 1997) and 246 sentences from Ba (2007). Conversely, the test data consist of 2364 unseen sentences randomly selected from Cissé (1994) and Ba (2007) which are disjoint from those sentences included in the development set. The average sentence length of the test set is 14.89 words; the longest sentence contains 70 words.

5. Experimental Evaluation

Evaluation of the Wolof LFG parser was conducted based on two metrics: coverage and accuracy. Coverage gives statistics about the test sentences that could receive one or more parses. To increase coverage, two robust XLE techniques were used: fragmenting and skimming. Fragments are produced when the grammar is unable to provide a full parse for the input sentence. This mechanism allows the parser to build for the input a sequence of well-formed chunks with both c-structure and f-structure associated with them. Likewise, skimmed parses are produced, when the amount of time or memory spent on a sentence exceeds a threshold, thereby avoiding timeout and memory problems.

Coverage is measured by breaking down the LFG scores according to whether the parser yields full parses or non-full parses (i.e. FRAGMENT, SKIMMED, or SKIMMED+FRAGMENT parses). Thus, coverage is defined as the percentage of parsed sentences in relation to the full corpus, and to measure it, an evaluation was conducted against the 2364 test sentences. As the results in Table 1 show, the parser could find a complete parse for 1712 of the test sentences (i.e. 72.72% coverage for complete parses). Conversely, 27.27% of the test sentences couldn't be parsed using the grammar alone. The use of the different robustness techniques allowed to increase the coverage to 99%.

The total and percentages of test examples in different classes of parse quality are listed in the first and second row of Table 1. The third and fourth row show the average score for the sentence length and parsing time to the respective classes. LFG scores broken down according to classes of parse quality are recorded in the next rows. The first column shows coverage scores for all parses in the test set. The second column shows the coverage scores when restricting attention to examples which receive only full parses. Columns 4-6 break down non-full parses according to examples which receive only FRAGMENT, only SKIMMED, or SKIMMED+FRAGMENT parses. As can be seen, unparsed or timed out sentences are mostly long sentences.

A weakness with the coverage metric is that it does not guarantee that the assigned parse is indeed the correct one (Carroll et al., 1998). Therefore, accuracy was used in addition to assess parsing quality. Accuracy is measured based on a detailed error analysis of the grammatical sentences which were incorrectly annotated by the parser. Due to a lack of a gold standard annotated corpus for Wolof, evaluation of accuracy was done using the set of full parses and skimmed sentences, excluding SKIMMED+FRAGMENT. However, since the parser usually produces a huge number of solutions, reviewing these by hand to see if the correct parse is in the output would be time-consuming.

Fortunately, the LFG Parsebanker (Rosén et al., 2009) from the INESS platform[4] provides an efficient and elegant solution for this issue. This web-based toolkit facilitates parse disambiguation by means of (lexical, morphological, c-structure and f-structure) discriminants. A discriminant can be defined as "any local property of a c-structure or f-structure that not all analyses share" (Rosén et al., 2005, p. 380). Any given discriminant can induce a binary partition on the choice space. The selection of a discriminant (or its complement) amounts to the selection of one of the two partition elements, reducing the choice space accordingly.

In INESS, the parse results along with the discriminants can be visualized through the XLE-Web interface, allowing for quickly choosing the desired solution — if contained in the output. Thus, accuracy was measured using the two following criteria: (i) GOLD: perfect parse(s);[5] and (ii) NO GOOD: the correct parse was not among the choices.

With these criteria, assessment of the parsing quality is done using F-score defined as the harmonic mean of precision and recall ($f = \frac{2 \cdot precision \cdot recall}{precision + recall}$). We may note in passing that if attention is paid to the full parses only, precision and recall might not be an issue for manual evaluation. However, if we take the entire test corpus into account, we can define recall as the percentage of correct complete parses in relation to the **full corpus**; and precision as the percentage of correct parses in relation to the **set of complete parses**. The evaluation of the parser accuracy indicates 67.2% recall, 92.8% precision and an f-score of 77.9%. The evaluation of the 1712 sentences that get a full parse reveals that, for 125 sentences, the correct parse was not among the choices, i.e. 10% were marked "NO GOOD". However, about 90% of the sentences that received a complete parse passed the accuracy test (and therefore marked "GOLD").

6. Discussion

6.1. Coverage

Among the well-formed sentences which received a partial parse, three types were distinguished: (i) constructions which could be handled by the grammar, but get FRAGMENT due to skimming techniques; it turned out that ca. 60 of such constructions were affected by this problem; (ii) constructions for which the grammar does not have rules, e.g. certain types of non-constituent coordination, certain parenthetical and VP ellipsis constructions; and (iii) sentences which contain lexical material that is not in the lexicon (e.g. subcategorization problems) or not covered by the morphology or the tokenizer (e.g. foreign language material and multiword expressions).

6.2. Parsing Quality

A detailed error analysis was performed on the grammatical sentences which were incorrectly annotated by the parser, as shown in Table 2. All no good sentences are reviewed by hand to check the c- and f-structures of the analyses. Subsequently, a score is assigned according to the number and (sub)type of errors.

In this evaluation, most relevant errors are ranked according to their frequency and classified into: wrong phrase structure

[4] See `http://iness.uib.no/iness/`.

[5] Some sentences will have more than one "correct parse".

	all	full	non-full			
			fragment	skimmed	skimmed + fragment	timed out
Items	2364	1712	294	15	333	10
% of test set	100%	72.72%	12.48%	0.63%	14.14%	0.42%
avg. sent. len	14.89	11.97	13.98	25.6	29.83	28.8
avg. time (CPU sec)	2.77	0.46	1.0	9.61	12.84	104.83

Table 1: LFG scores for the Wolof test examples

Rank	Total	%	Error Type	Error subtype	Subtotal
#1	63	51	Wrong PS (non-embedded NP)	Coordination	28
				Parentheticals and appositives	13
				Focus and copula constructions	13
				Bare infinitive	12
				Ellipsis	6
#2	24	20	Wrong PS (embedded NP)	Bound relative clauses	7
				Free relative clauses	7
				Bound and free relatives mismatch	6
				Interrogative nominal CP	4
#3	17	14	MWE	-	17
#4	8	8	Pronominal reference	-	8
#5	3	2	PP attachment	-	3
#6	1	1	Missing entry	-	1
#7	5	4	Misc	-	5

Table 2: Results of the corpus-based error analysis

(PS), excluding NPs and CPs with a nominal distribution; wrong PS in embedded NPs and nominal CPs; multiword expressions; pronominal reference; PP attachment; missing lexical entry; and other diverse errors (Misc).

The category #1 involves cases where the main clause got the wrong phrase structure because the parser made a mistake by assigning the wrong POS or subcategorization frame given the context. This error type mainly includes wrong coordination (mostly due to the lack of an overt conjunction), errors related to the treatment of appositives and parentheticals, wrong focus and copula constructions and bare infinitives.

Category #2 refers to a wrong phrase structure in a nominal embedded clause (NPs and CPs with a nominal distribution). It includes incorrect analyses of bound / free relatives or a mismatch of both clause types, as well as wrong interrogative CPs which distribute like nominals.

As Table 2 shows, mistakes made by the system in assigning the wrong phrase structure contributed the largest number of errors: 71%. In particular, some notorious issues were caused by the morphological ambiguity between bound/free relative pronouns and interrogative pronouns and their mismatch with determiners and adverbials. Also, coordination is one of the hotspots of ambiguity that leads to a large amount of incorrectly annotated phrase structures.

Moreover, many constructions were falsely identified as copula and focus related clauses due to an important number of mismatches between infinitival complementizers, relative pronouns, and demonstrative determiners. Besides, some sentences couldn't be parsed because they contain syntactic constructions for which the grammar does not have a rule or the rule was turned off for efficiency reasons. This includes e.g. non-constituent coordination, and VP ellipses.

7. Conclusion

This paper has discussed the implementation of the first Wolof parsing system based on the LFG formalism. The discussion highlighted various key issues in this language, including the treatment of noun classes, cleft and copula constructions, coordination, and some valency changing phenomena (causative and applicative). It has also shown that the implementation of these issues starts as early as tokenization, and goes through morphological and syntactic analyses.

Evaluation of the parsing system is done by running test suites of natural data through the grammar. This resulted in full grammar coverage on 2364 test data when combining robust parsing strategies with partial parsing techniques. A semi-automatic, discriminants-based approach allowed for disambiguating the parse output in an efficient way and for building a treebank that contains the maximum correct analysis or analyses possible. It has also allowed for evaluating accuracy of the parser. The results according to parse quality show that the full parses achieve more than 90% accuracy and a high f-score. These results are roughly comparable to those reported for other languages within ParGram such as Japanese (Masuichi et al., 2003).

However, the parsing evaluation reveals that a number of the full parses are not analyzed properly due to problems related to tokenization, incompleteness of the morphology, the use of robust parsing techniques, ambiguity and computationally expensive constructions. These findings have a number of important implications for strategies to control ambiguity and increase parsing efficiency.

8. Bibliographical References

Attia, M. (2008). Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation. Ph.D. thesis, University of Manchester.

Ba, M. (2007). Bataaxal bu gudde nii. Nouvelles Editions Africaines du Sénégal (NEAS).

Beesley, K. R. and Karttunen, L. (2003). Finite State Morphology. Center for the Study of Language and Information, Stanford, CA.

Bresnan, J. and Moshi, L. (1990). Object Asymmetries in Comparative Bantu Syntax. Linguistic inquiry, 21(2):147–185.

Bresnan, J. (1982). The passive in lexical theory. In Joan Bresnan, editor, The Mental Representation of Grammatical Relations, pages 3–86. The MIT Press, Cambridge, MA.

Bresnan, J. (2001). Lexical-Functional Syntax. Blackwell, Oxford.

Butt, M., King, T. H., Niño, M.-E., and Segond, F. (1999). A Grammar Writer's Cookbook. CSLI, Stanford, CA.

Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). The Parallel Grammar Project. In Proceedings of the COLING02 Workshop on Grammar Engineering and Evaluation, volume 15, pages 1–7. Association for Computational Linguistics.

Butt, M. (1995). The structure of complex predicates in Urdu. CSLI Publications, Stanford, CA.

Carroll, J., Briscoe, T., and Sanfilippo, A. (1998). Parser evaluation: a survey and a new proposal. In Proceedings of the 1st International Conference on Language Resources and Evaluation, pages 447–454.

Cissé, M. (1994). Contes wolof modernes. L'harmattan.

Comrie, B. (1985). Causative verb formation and other verb-deriving morphology. Language typology and syntactic description, 3:309–348.

Crouch, D., Dalrymple, M., Kaplan, R., King, T., Maxwell, J., and Newman, P. (2019). XLE Documentation. Online documentation, Palo Alto Research Center (PARC).

Dalrymple, M., Dyvik, H., and King, T. H. (2004). Copular Complements: Closed or Open? In Miriam Butt et al., editors, The Proceedings of the LFG '04 Conference, University of Canterbury.

Dalrymple, M., King, T. H., and Sadler, L. (2009). Indeterminacy by underspecification. Journal of Linguistics, 45(01):31–68.

Dione, C. B. (2012a). A Morphological Analyzer For Wolof Using Finite-State Techniques. In Proceedings of the 8th LREC. ELRA.

Dione, C. B. (2012b). An LFG Approach to Wolof Cleft Constructions. In Miriam Butt et al., editors, The Proceedings of the LFG '12 Conference, Stanford, CA. CSLI Publications.

Dione, C. B. (2013a). Handling Wolof Clitics in LFG. In Christine Meklenborg Salvesen et al., editors, Challenging Clitics, Amsterdam. John Benjamins Publishing Company.

Dione, C. B. (2013b). Valency Change and Complex Predicates in Wolof: An LFG Account. In Miriam Butt et al., editors, The Proceedings of the LFG '13 Conference, Stanford, CA. CSLI Publications.

Dione, C. B. (2014). LFG parse disambiguation for Wolof. Journal of Language Modelling, 2(1):105–165.

Dione, C. B. (2017). Finite-State Tokenization for a Deep Wolof LFG Grammar. Bergen Language and Linguistics Studies, 8(1).

Frank, A. (2002). A (discourse) functional analysis of asymmetric coordination. In Miriam Butt et al., editors, The Proceedings of the LFG '02 Conference, National Technical University of Athens.

Nataali Dominik Garros, editor. (1997). Bukkeek "perigam" bu xonq: teeñ yi. Dakar: SIL; Paris: EDICEF.

Kaplan, R. M. and Maxwell, J. T. (1988). Constituent coordination in Lexical-Functional Grammar. In Proceedings of the 12th conference on Computational linguistics-Volume 1, pages 303–305.

Kaplan, R. and Newman, P. (1997). Lexical resource reconciliation in the Xerox Linguistic Environment. In Proceedings of the ACL Workshop on Computational environments for Grammar Development and Engineering.

Kaplan, R. M., Maxwell III, J. T., King, T. H., and Crouch, R. (2004). Integrating Finite-State Technology with Deep LFG Grammars. In Proceedings of the ESSLLI'04 Workshop on Combining Shallow and Deep Processing for NLP.

Masuichi, H., Okuma, T., Yoshimura, H., and Harada, Y. (2003). Japanese parser on the basis of the Lexical-Functional Grammar formalism and its evaluation. In Proceedings of PACLIC17), pages 298–309.

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

McLaughlin, F. (2004). Is there an adjective class in Wolof? In R.M.W. Dixon et al., editors, Adjective classes. A crosslinguistic typology., pages 242–262. Oxford University Press.

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

Rosén, V., Meurer, P., and De Smedt, K. (2005). Constructing a parsed corpus with a large LFG grammar. In Miriam Butt et al., editors, Proceedings of the LFG'05 Conference, Stanford, CA. CSLI Publications.

Rosén, V., Meurer, P., and de Smedt, K. (2009). LFG Parsebanker: A Toolkit for Building and Searching a Treebank as a Parsed Corpus. In Frank Van Eynde, et al., editors, Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT7), pages 127–133, Utrecht. LOT.

Sulger, S., Butt, M., King, T. H., Meurer, P., Laczkó, T., Rákosi, G., Dione, C. B., Dyvik, H., Rosén, V., De Smedt, K., Patejuk, A., Çetinoğlu, O., Arka, I. W., and Mistica, M. (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.