International Journal of Computer Engineering Science (IJCES)
Volume 2 Issue 4 (May 2012) ISSN : 2250:3439
https://sites.google.com/site/ijcesjournal http://www.ijces.com/

Formalization of the Wolof with NooJ: Implementation on the Wolof dictionary

Haby Diallo, Alex Corenthin, Claude Lishou

Laboratoire de Traitement de l'Information
Ecole Supérieure Polytechnique-UCAD-BP 5085 Dakar-Fann, Sénégal
{haby42, claudelishou}@yahoo.fr, alex.corenthin@gmail.com

Abstract. This paper introduces the NooJ module for the Wolof language and the implemented electronic dictionary. the linguistic resources used are common usage dictionaries including Arame Fall's and Jean Léopold Diouf's dictionaries, as well as the Wolof lexicon available at CLAD (Centre de linguistique Appliquée de Dakar).

The present work will first focus on the socio-linguistic situation of the Wolof language while describing its alphabet before, in a second part, introducing the complex morphology of this language. The third part will be devoted to explaining how the core of the dictionary has been constructed and describing the flexional and derivation rules used to implement it in NooJ. Finally, the first results achieved with NooJ will be presented.

Keywords: Wolof, dictionary, NooJ, parsing, derivational morphology, flexional morphology.

1 The Wolof language

1.1 Socio-linguistic situation

Wolof belongs to the Atlantic group of the Niger-Congo family. Within the group, it had initially been classified in a North sub-group, with the Sereer and Ful, but recent studies show that Wolof has much in common with the ñuñ set, from a tenda-ñuñ sub-group. Wolof, a relatively homogeneous language, although with regional, particularly lexical, varieties, is spoken and understood by over 80% of Senegalese. In 1999, the number of Wolof speakers in Senegal was estimated to be higher than seven million. But it is the mother tongue of 44% of the total population of the country. "The Wolof" wrote Malherbe, "is one of the African languages whose cultural expansion is unquestionable: it (the language) is increasingly becoming and every day more the language of communication among Senegalese of different ethnic groups." It is part of autochthonous languages of Senegal which number twenty and are called the status of national languages as they are codified. In importance order, the six languages called of the first generation, can be listed. Wolof is spoken by 70.9% of the Senegalese population, according to official figures of the 1988 general census, Pulaar (21.1%), Sereer (13.7%), Manding (6.2%), Jola (5.2%), and Sononke (1.4%). These languages were officially provided with an alphabet as early as the first independence years, precisely in 1968.

They are used in the alphabetization and some of them in the experiments performed in formal education.

1.2 Wolof alphabet

The official Wolof alphabet counts twenty- seven (27) letters, twenty one (21) of which consonants and six (6) vowels that appear all in the Unicode standard (Unicode). These letters are the following:
a, b, c, d, e, ë, f, g, i, j, k, l, m, n, ñ, ŋ, o, ó, p, q, r, s t, u, w, x, y.

1.3 The phonetic

- In Wolof, all the consonants can be geminated with the exception of *f, h, q, s* and *x*. Geminating is relevant and appears in internal position and end of lexemes. It is marked by the reduplication of the consonant.
  Example:

Table 1. Reduplication of the consonant

Word	Translation	Word with consonant duplicated	Translation
lemi	go to the bend	lemmi	unfold
bët	eye	bëtt	pass across
sëg	cemetery	sëgg	to stoop, to bend

- All the occlusive consonants of Wolof can be penalized. To spell them, the nasal **m** is used before the labials **b** and **p**, and the nasal **n** before all other consonants:

Table 2. Penalized occlusive consonants

Consonnant	Word	Translation
Mb	mbóot	cockroach
Mp	nàmp	suckle
Nq	xonq	red
Nt	bunt	door

- The voiced occlusives are performed mute in the absolute final of lexeme. But they are always recorded in their voiced form, when writing.

Table 3. Voiceless occlusive consonants

Word	Translation	Word	Translation
fab	take	fabul	he did not take
soj	rheum, to be cold	sojul	he is not cold
néeg	room	néegam	her room
toob	last name	toobeen	toob family

- The graph ë/Ë is chosen to spell the overage middle vowel.

ë =>

bënn “ to pierce”
xëcc “to draw”
xëm “to faint, to chars”
ëpp “ to be too, to fan”

Fig. 1. Graph ë/Ë

- For **e** and **o**, there is a relevant aperture opposition. The closure is marked by the acute accent.

fel “feal” /fél “strike”
xel “intelligence” /xél “leak, speed”
jodd “straight” /jódd “typical dish made from cassava”
xott “shell” /xótt “onomatopoeia expressing effeminacy

- There are three kinds of a vowels in Wolof:
  - vowel written **a/A**
  - on open vowel written **à/À**
  - a nasal vowel **ã/Ã**

Table 4. Example of using the vowel a

Word	Translation
am	to have
àll	bush
ãhãa!	oh well

- In the Wolof vowel system, to each short vowel corresponds long one, with the exception of the vowel /**à**/**:**

Table 5. Vowels Wolof

The brief			Long		
Vowel	Word	Translation	Vowel	Word	Translation
a	gal	pure silver/ to swell(foot)	Aa	gaa	canoe
à	àll	bush	…	…	
ã	ãhã	yes	àa	ãhãa !	oh well !
e	set	clean	ee	seet	to look for
é	wér	healthy	ée	wéer	to lean
ë	bër	vacancy	ëe	bëer	variety of fish
i	bir	to be clear / audible	ii	biir	stomach
o	xol	heart	oo	xool	to watch
ó	xótt	onomatopoeia expressing the softness	óo	xóot	deep
u	tur	first name	uu	tuur	to shed

- When the long vowel (sequence of identical ones) is stressed, only the first vowel carries the accent. It is the same with the central vowel.

Table 6. Use the long vowel

Word	Translation
néeg	room
xóot	deep
beer	Variety of fish

Due to vowel harmony, when the first vowel of a word is closed (I,u,e,o), all the following syllable vowels are pronounced closed. By virtue of this predictability, only the first closed vowels are marked with an accent, when it comes to **e** or **o**.

Table 7. Example of use

Word	Phonetic	Translation
ndim**o**	[ndim**o**]	tissue
pus**o**	[pus**o**]	needle
gor**e**	[go:r**e**]	to human
téer**e**	[te:r**e**]	book
su wut**e e**	[su wut**e:**]	if he seeks

Wolof is of a great vitality in oral communication; it is widely used in all daily life action, in the audio-visual press, in religious sermons among Moslems as well as among Christians, in advertizing etc…

1.4 Computer tools Wolof

Wolof is present on the web, but no norms are observed concerning the way characters are written on the pages encountered on the web or digital documents. At the beginning the aim was to promote the language on the web, to make it known without worrying about writing. For the language self-starting processing purposes the writing has to be standardized and to abide by the codification. The lack of specific tools for Wolof TALN, such as keyboards for entering characters and spell checkers, slows down this effort. However, thanks to keyboards such as AFRO [6] developed to treat the characters of African languages, it is possible to write Wolof. For example, combining "**n**" and "<" makes it possible to write the character "ŋ".

Today, several standards are designed to write languages that are not taken into account by the set of Latin characters [7]. Unicode is part of the standards which provide a chance to African languages to be coded by using a universal standard. So, Unicode brings an answer to the coding problem with its enlarged character set (over 2 bytes). The UTF-8 encoding even processes ASCII characters: Each character, whether or not ASCII, has a unique code assigned to it.

1.5 Lexical categories

As for the lexical form, in Wolof it is possible to distinguish:

- monovalent lexemes: they are nominal lexemes in all the uses they can have in the language, or always verbal function lexemes.

Example:

- nominal lexemes
  nit: human
  sabar : drum
- verbal lexemes :
  xeeñ : smell
  dagg : cut

- Bivalent lexemes: they are part of no precise category. They are available for a nominal function as well as for verbal function.

Table 8. Example of bivalent lexemes

Word	Verbal function	Nominal function
dem (to leave)	den na (he has left)	dem bi (departure)
dof (mad)	dof na (he has gone mad)	dof bi (the mad person)

- The derivation: In this category there are lexemes from simple derivation and from complex derivation. The derivation rules will be examined below in another part of this paper.
- Composition: it is the result of the combination of the syntagm combination of forms which, in principle, are all lexemes. The commonest procedure encountered here consists in combining similar forms.

1.6 Lexicographical resources

There are easy to use dictionaries where each entry is followed by some information indicating a noun, or a verb or about the translation into another language. These include the Wolof-French and French-Wolof [5], The Wolof-French dictionary followed by a French-Wolof index [8]. Beside these dictionaries, it is to mention the basic Wolof, an alphabetical and analytical lexicon of Wolof in five volumes of Dakar Applied Linguistics Center (CLAD). But also mention the vocabulary of the fauna in Senegal [4] and the vocabulary of the flora in Senegal [4] and the Wolof-French vocabulary elections [1]. There are also on line resources they include among LEXICOLOG[1], Afro Web[2]. These resources however, are not in harmony with the structure used the NooJ dictionaries: They are in the lemma/information form (definition or translation into another language such as French). These resources were used in this work to gather the inflected lemmas on which inflexion and derivation rules were applied to thus automatically generate the inflected forms. This process is by taking each word back to its basic form to make it an electronic dictionary adaptable to NooJ.

For the Wolof language, as well as other languages in Senegal, descriptions are available which generally have been made as part of university theses or dissertations in Senegal, in States sharing the same language or in American and European universities. These studies provide a survey on the concerned language phonology, lexicology and grammatical structure. They have made it possible to carry out the applied research work required by the concrete use of national languages in the modern sector, particularly in education. This research work includes, high on the list, general dictionaries [8] and [5], terminology dictionaries [3], orthographical dictionaries, practical grammars [2], and teaching manuals in different fields etc… It is worth noting the lack of research work in automatization of these tasks. For example, there are no grammatical, spelling specific correctors which can process the writing of this language. This language NooJ formalism will be first step towards its automatic treatment.

[1] http://www.lexilogos.com/wolof_dictionnaire.htm

[2] http://afroweb.chez.com/frm_wofr.htm

2 Description of the Wolof dictionary

Each entry into the Wolof dictionary generally presents the following details:

- the lemma,
- the lexical category,
- French translation,
- the gender (for nouns varying in gender),
- the singular form (for nouns not varying in gender).

2.1 The grammatical categories

In the Wolof dictionary, each entry is unambiguously associated with a grammatical category designated by a written code with capital letters. The codes are listed in the following table:

Table 9. Grammatical category

Category	Code
Adjective	ADJ
Noun	N
Verb	V
Adverb	ADV
Preposition	PREP
Pronoun	PRON
Demonstrative	DEM
Relative	REL
Determinant	DET

In addition, optional parts are associated with entries into the dictionary:

- Syntactic-semantic information: Most of the codes have assigned to nominal lexical entries. The features are as follows:

Table 10. Semantic features

Syntactic-semantic features	code	Examples
Abstract	Abs	Jàmm ji (the peace)
Concrete	Conc	Sër (loin cloth)
Animal	Amim	Lëpp-lëpp (butterfly)
Vegetal	Veg	Bissap (sorrel)
Human	Hum	Astu (Astou)
Medical	Medic	Jarag ji (the patient)
Places/Location	Loc	Ndakaru (Dakar)
Date/Hour	Date	Weru koor(the Ramadan month)
Organization	Org	Tostan

- a flexional paradigms call introduced by “+FLEX”

- a derivational paradigms call introduced by "+DRV"

2.1 Morphology

2.1.1 Verbal morphology

In Wolof, there are transitive verbs, intransitive verbs and auxiliary verbs. These verbs are conjugated in different types and modes. In this case, two aspects are shown. These are:

- Accomplished: affirmative present or negative present, affirmative past or negative past;
- Unaccomplished: this affirmative or negative present, past, affirmative or negative past.

The inflection and derivation rules for verbs will be described below in another part of the present work.

2.1.2 Nominal morphology

In this part, it is possible to distinguish:

- Simple names: they have the following characteristics:
  - the kind
  - the number

Examples: Askan, N + s + FR=race (EN= race)

- Compound nouns: these are nouns consisting of basic lemmas. The following compositions can be noted:
  - the same form repeated
    Example:
    jam: to prick and jam-jam : prick
  - combination of forms
    Example:
    jaam: slave and buur: King
    yield jaambur: free person
  - several combined lexemes
    Example:
    jambuur: free person
    yield jambuur-jambuur: inhabitant of the land of the free people

2.1.3 Pronouns

For the pronouns, the following set was numbered:

- 18 subjects, objects and emphatic forms pronouns, emphatics and objects included
- 40 demonstrative pronouns
- 20 relative pronouns

The following tool words have been listed:

- 28 definite and indefinite articles included
- 10 interrogative
- 19 quantifying
- 25 adverbial
- 21numerals

2.1.4 Deverbals

They are nouns formed from the radical of verbs. In Wolof most of deverbals are built by consonantal gradation of the root initial consonant [Robert S.].

Example: sàcc: to steal yields càcc: theft

3 Automatic Wolof language processing

3.1 NooJ [12]

NooJ is a development linguistic environment which provides tools to build test and maintain wide coverage formalized description of natural languages (in the form of dictionaries and electronic grammars). The dictionaries and grammars are applied to texts in order to locate morphological, lexicological and syntactic models, remove ambiguities, and classify some simple and compound words.

NooJ is used as a development linguistic platform, a documentary research system, a terminology extractor, as well as to teach linguistics and computational linguistics to students [10] and [11].

3.2 The dictionary construction

Morphology rules have been expressed in the NooJ format.

3.2.1 Morphological rules

NooJ has at its disposal routines and algorithm that allow the genesis of the inflected forms set based on a call to inflectional and derivational descriptions assigned to each dictionary entry. These descriptions can be described in the shape of rational expressions or graphs established with the help of the NooJ graphic editors and the use of predefined generic controls.

In Wolof the rules that will be used to make it possible to do:

- The suffixation: add at the end of the "suffix" variable content. This word can be a verb or, a noun.

Table 11. Suffixation rule

The "suf" rule
suf=suf/V
suf=suf/N

Example: suf=waale
juddu (to be born)=> judduwaale (to be innate)

- The prefixation: adds at the beginning of the word the content of the "prefix" variable. Two cases are presented :a hyphen (-) is added after the prefix variable (cf. Table 11) to thus obtain a compound word in the first case and in the other one it will just be added at the beginning of the word(cf. Table 12).

Table 12. First prefixation rule

The "prefixe" rule
prefixe=<LW>prefixe-/N

Tableau 13. Second prefixation rule

The "prefixe" rule
prefixe=<LW>prefixe/V

- The root initial consonant gradation: to replace the first character by the content of the "alter" variable

Table 14. Rule of alternance

Rule "alter"
alter=<LW><S>alter/N

Example: fo (play) => po (game)

- The combination of rule: consonantal gradation and suffixation; to add at the end of the word the "suffixe" variable content and the go to the beginning of the word to replace the first character by the "alter" variable content.

Table 15. Gradation and suffixation

The “altersuffixe” rule
altersuffixe =suffixe<LW><S>alter<>/N

Example: damp (to massage) => ndampaay (massage)

3.2.2 Nouns

As for nouns, there are nouns with forms that do not vary when they are plural and nouns that change their forms when plural. To make a gender distinction, the phrases “bu góor” (for the masculine) and “bu jiggèen” (for the feminine) are used. But however some exceptions exist concerning animal names where the spelling of the noun changes when passing from the masculine to the feminine. Automatons have been defined for these nouns to ensure the gender and number inflexion rules.
Example: the transformation of bët(eye) in gët(eyes).

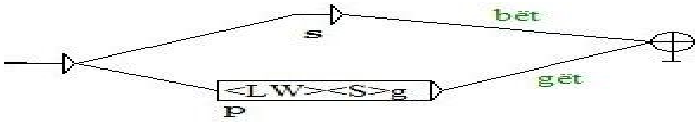

Fig. 2: Flexional grammar of the word “bët”

3.2.3 Derivation rules of verbs

For verbs in Wolof, inflexion rules according to the tense, the aspect and the modality used follow. Sixty seven (67) forms for conjugation have been listed. But only the following eight have been short listed with the use of the verb “bey” (cultivate):

- APMF (Accomplished-Present-Amodal-Affirmative)

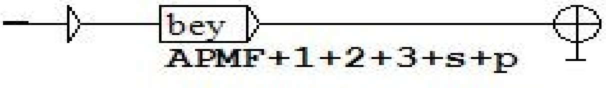

Fig. 3 : The APMF grammar

- APMN (Accomplished-Present-Amodal-Negative)

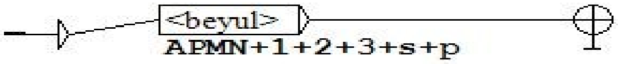

Fig. 4 : The APMN grammar

- ASMF (Accomplished-Past-Amodal-Affirmative)

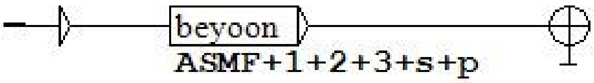

Fig.5: The ASMF grammar

- ASMN (Accomplished-Past-Amodal-Negative)

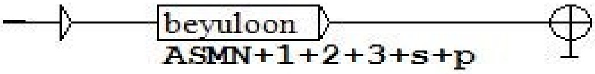

Fig.6: The ASMN grammar

- APON (Accomplished-Present-Circonstant-Negative)

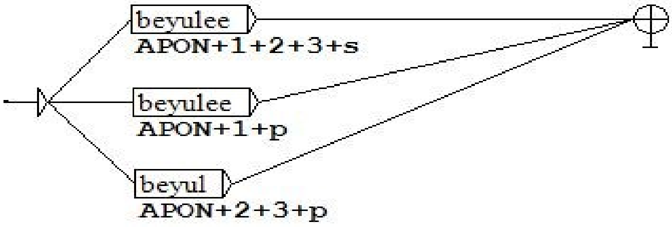

Fig.7: The APON grammar

- ASVN (Accomplished- Past- verb emphasized -Negative)

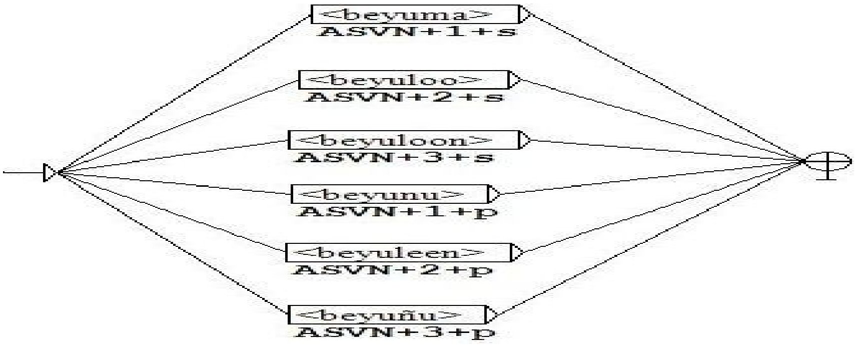

Fig.8 The ASVN grammar

- APOF(Accomplished-Present-circonstant-Affirmative)

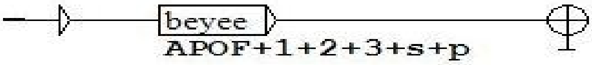

Fig. 9 : The APOF grammar

- APVN (Accomplished- Present-verb emphasized-Negative)

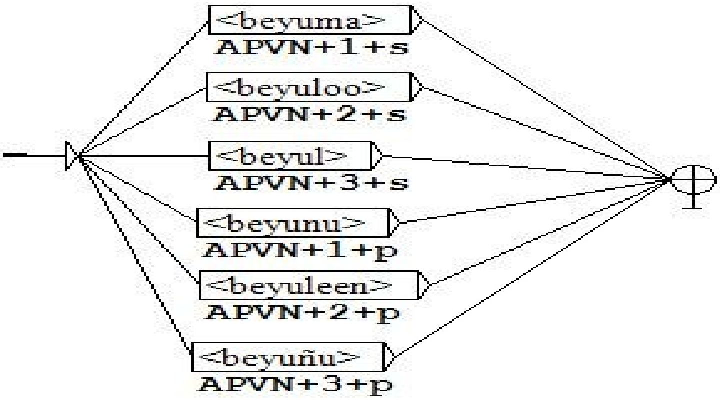

Fig.10: The APVN grammar

- APTF(Accomplished-present-imperative-negative)

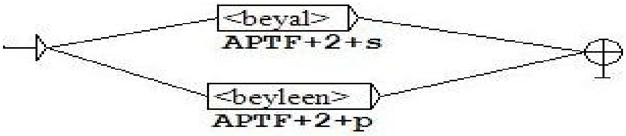

Fig.11: The APTF grammar

Ten inflexion models have also been listed for all the verbs. The following model BEY is given as an example:

```
  BEY=<E>/INF
     + <E>/APMF+1+2+s + <E>/APMF+1+2+P
     + ul/APMN+1+2+s + ul/APMN+1+2+p
     + oon/ASMF+1+2+3+s + oon/ASMF+1+2+3+p
     + uloon/ASMN+1+2+3+s + uloon/ASMN+1+2+3+p
     +  uma/APVN+1+s  +  uloo/APVN+2+s  +  ul/APVN+3+s  +
  unu/APVN+1+p  + uleen/APVN+2+p + uñu/APVN+3+p
     + uma/ASVN+1+s + uloo/ASVN+2+s + uloon/ASVN+3+s +
unu/ASVN+1+p + uleen/ASVN+2+p + uñu/ASVN+3+p
     + ee/APOF+1+2+3+s + ee/APOF+1+2+3+P
     + ulee/APON+1+2+3+s + ulee/APON+1+p + ul/APON+2+3+p
     + al/APTF + leen/APTF
     +<E>/APTN+1+s + <E>/APTN+1+p
```

3.2.4 Verbal entries

For verbal entries, each verb will be carried back to the third person singular at the accomplished present amodal affirmative, then associated with the flexion model chose in the present work, then associated again with one or several derivation models among the thirteen models developed for all verbs. A verb entry is of the form:

Lekk, V + T + + DRV FLX = BEY = Antu

In addition to the symbol designating the grammar category ("V" for the verbs), each entry is supplied with the information linked to its transitivity.

The expression "T+" indicates that the verb used is transitive.

The expression "+FLX=BEY" indicates a call to the BEY flexional paradigm to use to inflect the entry lemma. This flexional paradigm designates the information set that makes it possible, starting from the lexical entry, to automatically generate the set of its conjugated form. Wolof verbs have an average of forty four inflected forms.

The expression "DRV + = Antu" designates an "Antu" derivation paradigm call to use.

To enable the present module to go smoothly a few grammars have been designed to recognize days, month and years in Wolof as well as grammar enabling some simple phrases and sentences to be translated into French.

A few grammars have also been built that make it possible to recognize Wolof phrase. These grammars include the one that recognizes a date in Wolof (cf. Fig. 12).

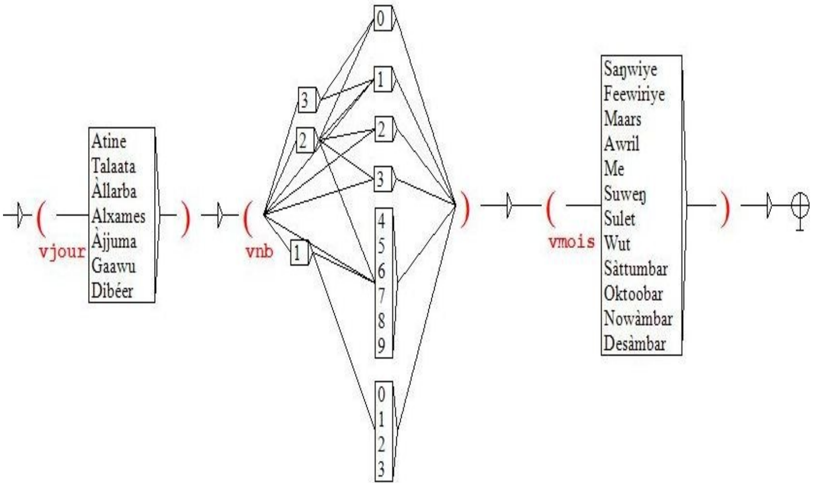

Fig. 12 : Grammar of dates in Wolof

A grammar has thus been created which contains phrases making it possible to ask what time it is in Wolof (cf. Fig. 13) followed by their translations into French.

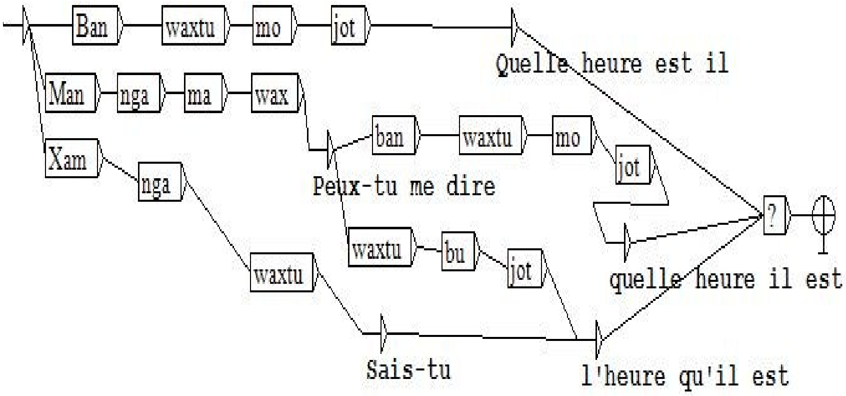

Fig.13: Grammar of the expression of time it is in Wolof

3.2.5 Results

Two Wolof dictionaries have been built for NooJ. One dictionary containing 2562 verbs and another one containing 8567 entries for simple nouns and compound nouns but also taking into account the pronouns and numerals.

The dictionary designed within the framework of the present research work has been tested on the text of the Universal Declaration of Human Rights in Wolof which contains 577 different words. After analysis, three hundred and thirty two (332) words or 58.06% are recognized. This is due to problems related to the spelling of some words. But it can also be due to errors generate during the dictionary. To bring a remedy, either the no recognize words have to be introduced into the dictionary, or a spelling corrector has to be produces in order to first correct the texts and text corpus before a possible analysis.

For a first analysis and before, for the moment, an unfinished dictionary the result achieved can be judged partly satisfactory.

4 Conclusion

The aim of the present search work was to produce and formalize Wolof with NooJ by producing dictionaries and grammars in this language. To finalize this first attempt the designed dictionaries grammars have to be improved and NooJ other features such as the recognition of named entities in Wolof and the development of analyzers and spell checkers have to be developed for the same language while providing solutions appropriate for Wolof writing uniformity.

References

1. Diallo, A., Mbodj, C., Seck, A.N., Thiam, N. : Vocabulaire des élections wolof-français suivi d'un index français wolof, CLAD, Dakar, (1997)
2. Diagne, P. : Grammaire du wolof moderne, Présence Africaine, Paris, (1971)
3. Diaw, A.A. : Vocabulaire de la faune au Sénégal, CLAD, Dakar, (1976)
4. Diaw, A.A. : Vocabulaire de la flore au Sénégal, CLAD, Dakar, (1981)
5. Diouf, J.L. : Dictionnaire wolof-français et français-wolof, Edition Karthala, (2003)
6. Enguehard, C., Mbodj, C. : Des correcteurs orthographiques pour les langues africaines, BULAG, n° 29, (Centre tesnière), pp. 51-68 (2004)
7. Chanard, C., Popescu, B.A. : Encodage informatique multilingue : application au contexte du Niger. Les cahiers du RIFAL, n°22, pp.33-45 (2001) http://andreipb.free.fr/textes/Chanard-Popescu-2002.pdf
8. Fall, A., Santos, R., Doneux,J.L. : Dictionnaire wolof-français suivi d'un index français-wolof, édition Kartala, Paris, (1990)
9. Robert S. : Le Wolof, (sous presse) in Djamel Kouloughli & Alain Peyraube (éds), Dictionnaire des Langues, vol. 3 de l'Encyclopédie des Sciences du Langage, Sylvain Auroux (éd.), Paris : P.U.F
10. Silberztein, M. : Le dictionnaire DELAC, in Dictionnaires électroniques du français, Langue française n° 87, Paris, Larousse, (1990)
11. Silberztein, M., Tutin, A. : NooJ : Un outil TAL de corpus pour l'enseignement des langues et la linguistique, journées ATALA "TAL et apprentissage des langues",(2004)

12. Silberztein, M. ; Formaliser les langues avec l'ordinateur, Cahiers de la MSH Ledoux, Presses universitaires de Franche-Comté, (2007)