# Vers un analyseur syntaxique du wolof

*Mar Ndiaye*[1] *Cherif Mbodj*[2]
(1) Ecole supérieure de commerce Dakar, 7 av. Faidherbe BP21354 - Dakar
(2) Centre de linguistique appliquée de Dakar (UCAD)
ndiaye.mar@gmail.com, cmbodj@ucad.sn

RESUME

Dans cet article nous présentons notre projet d'analyseur syntaxique du wolof, une langue parlée au Sénégal, en Mauritanie et en Gambie. Le modèle d'analyse que nous utilisons est très largement inspiré du modèle d'analyse syntaxique multilingue de Fips (Laenzlinger et Wehrli, 1991 ; Wehrli, 1997,2004)[1] développé au LATL[2] de l'université de Genève, sur la base de grammaires inspirées des théories chomskyennes, notamment la grammaire GB[3].

ABSTRACT

**a futur syntactic parser for wolof**

This paper presents our project to implement a parser for wolof. The Wolof is an african language spoken in Senegal, Mauritania and Gambia. The project aims to implement a parser based to the Fips's grammatical model, a GB parser.

MOTS-CLES : wolof, TALN, analyseur syntaxique, Fips, GB.
KEYWORDS : wolof, NLP, syntactic parser, Fips, GB.

[1] Ce papier s'en est largement inpiré

[2] Laboratoire d'analyse et des technologies du langage

[3] *Government and Binding*

## 1 Introduction

Dans cet article, nous présentons l'architecture informatique du système. Cette architecture est entièrement basée sur celle de Fips. Ce choix est justifié par le fait que Fips utilise une technologie multilingue reconnue. Nous présentons d'abord rapidement le modèle grammatical sous-jacent à l'analyse syntaxique (section 2), ensuite nous abordons la structure des données linguistiques (section 3) et enfin la stratégie d'analyse (section 4).

## 2 La grammaire GB

Une grammaire GB est définie comme un système de principes, qui ne varient pas d'une langue à l'autre et de paramètres qui tiennent compte des propriétés spécifiques à chaque langue. Ces principes sont organisés en sous-systèmes appelés des modules. Chaque sous-système s'occupe d'un processus ou d'un groupe de phénomènes linguistiques. La théorie X-barre définit la structure hiérarchique en constituants de la phrase (FIGURE 1), la théorie du gouvernement règle les relations structurales entre les constituants, la théorie thêta s'occupe de l'assignation des rôles thématiques aux arguments, la théorie des cas règle la distribution des groupes nominaux dans la phrase. La théorie du liage s'occupe de l'interprétation (co)référentielle des groupes nominaux. La théorie des chaînes gère la constitution des chaînes entre les éléments déplacés et leurs traces laissées dans leur position d'origine.

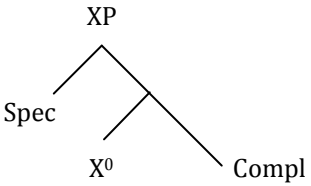



FIGURE 2 – Schéma X-barre

### 2.1 La théorie thématique

Cette théorie s'occupe de l'assignation des rôles thématiques aux arguments. Le prédicat verbal donne le corps de la phrase. Le verbe et ses arguments déterminent les constituants indispensables dans la phrase. Les relations sémantiques entre le prédicat et ses arguments sont spécifiées dans la grille thématique du verbe qui est une liste non ordonnée de rôles thématiques, dont les principales sont l'agent, le thème et le bénéficiaire.

### 2.2 La théorie du cas

Cette théorie s'occupe de l'assignation des cas aux syntagmes nominaux. Elle distingue deux types de cas: le cas structurel et le cas inhérent. Les cas structuraux sont assignés sous gouvernement de tête et comprennent le nominatif assigné par l'inflexion à son spécifieur,

l'accusatif, assigné par le verbe à son complément. Le cas inhérent est une propriété lexicale, c'est-à-dire un paramètre de la langue.

Pour satisfaire le filtre de cas, les syntagmes nominaux qui ne se trouvent pas dans une position où un cas peut être assigné peuvent se déplacer dans une position libre. C'est typiquement le cas du sujet, qui se déplace de la position spécificateur de VP, qui n'est pas une position de cas structurel à la position spécificateur de TP où il peut recevoir le cas nominatif ou encore le complément d'objet direct qui se déplace aussi en position spécificateur de TP dans les constructions passives à montée, car le verbe ne peut plus assigner le cas structurel à son complément.

### 2.3 La théorie des chaînes

Certains principes de la grammaire exigent que des éléments, projection maximale, tête, ne restent pas dans leur position canonique mais se déplacent dans d'autres positions. Le principe de projection et le principe de préservation de la structure exigent que la position de base continue d'exister, remplie par une trace de l'élément déplacé. Les mouvements sont codés dans des chaînes qui comportent les éléments déplacés et les traces qu'ils ont laissées dans leur position de base.

## 3 Le schéma X-barre dans Fips

Fips implémente une version simplifiée (FIGURE 2) du schéma X-barre standard de la théorie GB (FIGURE 1)

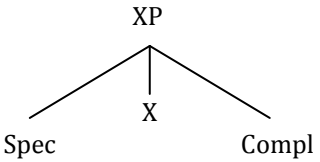



FIGURE 2 – Schéma X-barre dans Fips

La variable X (FIGURE 2), appelée tête, détermine la projection maximale XP. Elle prend ses valeurs dans l'ensemble constitué des catégories lexicales: Adv(adverbe), A(adjectif), N(nom), V(verbe), P(reposition) et fonctionnelles: C(omplementeur), Conj(onction), Interj(ection) et T(ense) (pour le morphème de temps/inflexion), D(eterminant) et F(onctionnel). Elle peut être modifiée par Spec et Compl qui sont des listes (éventuellement vides) de projections maximales correspondant respectivement aux sous-constituants gauches et droits de X.

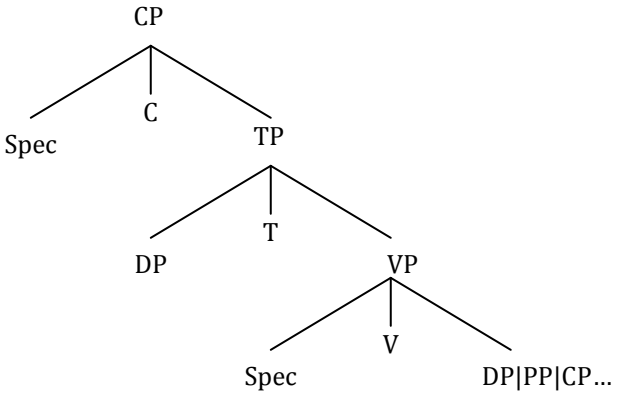



FIGURE 3 – Structure d'une phrase complète

Une phrase complète est représentée par une projection maximale de type CP (FIGURE 3). La catégorie C sélectionne une projection TP dans laquelle la position T comprend le verbe simple (ou l'auxiliaire) conjugué. Le sujet de la phrase est représenté au niveau Spec de TP alors que la position Compl de TP comprend le groupe verbal. La tête V du groupe verbal correspond à des verbes au participe passé ou des verbes à l'infinitif. La liste Spec de la projection VP est occupée par les adverbes(Adv) alors que Compl reçoit les autres arguments du verbe sauf le sujet. On doit à Pollock (1989) l'hypothèse de la montée des verbes conjugués de VP à TP - dans certaines langues (français, langues romanes) mais pas en anglais et dans les langues germaniques. Comme la tête T en anglais et dans les langues germaniques n'est pas suffisamment riche pour permettre la transmission des rôles thématiques portés par le verbe qui monterait s'y adjoindre à la trace de V - cette montée empêcherait donc la vérification du critère thématique -, expliquant ainsi pourquoi le verbe dans ces langues ne monte pas en T.

## 3.1 Le lexique

Le wolof est une langue morphologiquement riche. Par exemple, Voisin (20109) identifie les morphèmes - *i* et *si* comme encodant des valeurs telles que le mouvement associé, (exemples (1), (2), (3), (4)).

(1) a. dafa doon xataraayu nguir xeex-i

b. EV3S PASSE se. débattre pour se. battre-EL

c. *Il se débattait pour aller se battre*

(2) a. sa liggéey a ngi baax-si muñ-al tuuti rekk

b. POSS2S travail PRES3S ê.bon-RAPP patienter-IMP peu

c. *ton travail devient bon, patiente encore un peu.*

(3) a. Mu ngi ma-y nob-si

b. PRES3S O1S-INACC aimer-RAPP

c. il *(elle) devient amoureux (euse) de moi*

(4) a. Ndax ajuu na ñu seet-i ko

b. INTER ê.nécessaire P3S NAR1P regarder-EL O3S

c. *Est-il nécessaire que nous allons le voir*

La structure du lexique suit également le model lexical de Fips, c'est à dire un lexique relationnel selon lequel les relations morphologiques seront exprimées dans le lexique sous la forme de liens entre différentes représentations lexicales. Sans entrer dans les détails, la structure de la base de données lexicale s'articule comme suit (voir Seretan et *al.* , 2006) : nous avons (i) un lexique des mots, contenant toutes les formes fléchies des mots de la langue, ici le wolof, (ii) un lexique des lexèmes, contenant les informations syntaxiques de chaque unité lexicale (une unité lexicale correspond plus ou moins à une entrée de dictionnaire classique) .

Un exemple d'unité lexicale en wolof (tiré de (Mbodj et Enguehard, 2004) est donné en(3))

(3) forme : aay

phonétique : [a :y]

catégorie : v.i

mode de flexion : 2

définition : être mauvais, être mal

exemple d'usage: lu ayy ci li ma wax (qu'est-ce qu'il y a de mal dans ce que j'ai dit ?)

## 3.2 Le groupe nominal

Fips adopte l'hypothèse DP, selon la quelle la catégorie fonctionnelle D, réalisée comme déterminant, sélectionne un complément lexical NP à tête nominale. En d'autres termes, c'est le déterminant qui fonctionne comme tête du syntagme nominal.

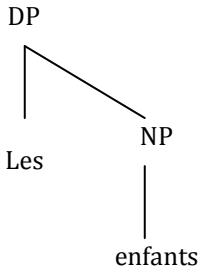


FIGURE 4 – Structure du groupe nominal

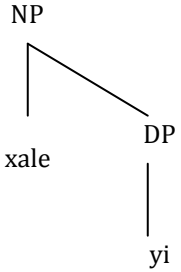
La structure du groupe nominal wolof est particulièrement intéressante dans le cadre de ce modèle d'analyse. Il s'avère qu'en wolof, le déterminant peut être en position post-nominale (exemple (4)) ou en position pré-nominale (exemple (5)). Ce qui nous oblige à reconsidérer la structure du DP adoptée.

(4) a. xale yi

b. enfants DEF.P

c. *les enfants*

d.

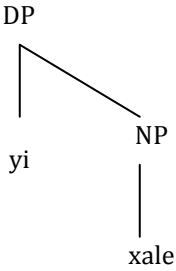



(5) a. yi xale

b. enfants DEF.P

c. *ces enfants*

d.




Dans l'exemple donné en (4) c'est le nom qui sélectionne un DP, alors que l'exemple (5) correspond à l'hypothèse adoptée dans Fips.

## 4 La stratégie d'analyse de Fips

La stratégie d'analyse de Fips (ALGORITHME 1) est de type gauche à droite avec traitement parallèle des alternatives. C'est une approche incrémentale essentiellement ascendante avec un filtre descendant. Les principes fondamentaux de l'algorithme 1 dit du "coin droit" sont:

- C'est une analyse syntaxique dirigée par les données. On cherche à attacher chaque nouvel élément au coin droit d'un constituant dans le contexte gauche.
- Le contexte gauche spécifie un ensemble de nœuds actifs auxquels le nouvel élément est susceptible de s'attacher (sites d'attachement).
- Tous les attachements possibles sont considérés en parallèle.

## 4.1 Type d'action

Fips utilise trois mécanismes fondamentaux qui sont : (i) la projection, (ii) la combinaison des constituants et (iii) le déplacement.

### 4.1.1 La projection

Le mécanisme de projection crée une structure syntaxique complète sur la base soit d'une structure lexicale, soit sur la base d'une structure syntaxique (par exemple un syntagme nominal à valeur adverbiale)

### 4.1.2 La combinaison

L'opération de combinaison implique deux projections adjacentes. Soient deux projections A et B, deux cas de figure se présentent:

- A est attaché comme sous-constituant gauche de B
- B est attaché comme sous-constituant droit de A ou d'un sous-constituant droit actif de A

### 4.1.3 Le déplacement

Dans la théorie chomskyenne, tout syntagme nominal qui n'a pas valeur d'adverbe doit être associé à un rôle thématique distribué par un prédicat sous condition de gouvernement. Les éléments extraposés sont des éléments déplacés par une transformation de mouvement à partir d'une position dite canonique, gouverné par un prédicat. Un syntagme nominal extraposé reçoit son rôle thématique par l'intermédiaire de cette position canonique à laquelle il reste lié (sous-section 2.3). Dans Fips, à un élément extraposé est associée une catégorie vide en position canonique d'argument (position sujet ou position complément). Le lien entre le syntagme nominal extraposé et le syntagme abstrait *e* qui représente sa trace en position canonique est établi par le même indice dans les deux structures

## 4.2 Exemple d'analyse

De façon très simpliste, sans entrer dans les détails de l'algorithme, nous allons montrer comment l'algorithme effectue l'analyse donnée en (7) pour la phrase donnée en (6).

(6) a. xale yi nelleewnañu

b. enfants DEF.P dormir

c. *les enfants dorment*

La lecture du premier mot de la phrase, *xale* donne lieu à une projection de type [NP xale]. Lorsque la tête de lecture lit le mot suivant, *yi* qui est un déterminant défini pluriel, l'action de créer crée une projection [DP yi]. Ce constituant est attaché comme sous-constituant droit

de NP, ce que donne le constituant [NP xale [DP yi]] (représenté en (4d.)). A la lecture du mot *nelleewnañu* , qui est un verbe conjugué, un projection de type [TP nelleewnañu [VP e]]. Cette dernière se combine avec le constituant [NP xale [DP yi]], attaché comme spécificateur de TP, c'est-à-dire comme sujet. Ce qui donne l structure arborescente (7) suivante:

(7)

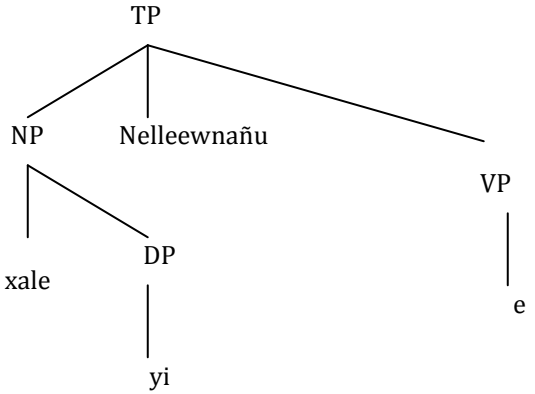



FIGURE 5 – Structure de la phrase *xale yi nelleewnañu*

## 5 Conclusion

La première phase du projet consiste à spécifier formellement la grammaire du wolof. Dans la deuxième phase, nous passons à la mise en œuvre informatique du lexique. La dernière phase concerne l'implémentation informatique de l'analyseur proprement dit sous *BlackBox Component Builder*, un système créé par *Oberon Microsystems Inc*. Le langage *Componenent pascal* est une extension du langage de programmation Oberon[4].

## Références

LAENZLINGER, C., WEHRLI, E. (1991). Fips un analyseur interactif pour le français *TA*

[4] Oberon est un descendant de Pascal et Modula-2 créé en 1985 par Niklaus Wirth et Jürg Gutknecht de ETH Zurich Switzerland (Wirth, 1985)

*informatiosn* , 32 :2, pages 35–49 .

MBODJ, C. et ENGUEHARD, C. (2004) Des correcteurs orthographiques pour les langues africaines. *BULAG* (bulletin de linguistique appliquée et générale), 29

POLLOCK, j.-Y (1989). Verb movement universal grammar, and the structure of IP. *LI,* 20(3) , pages 365-424.

SERETAN, V., WEHRLI, E. et NERIMA, L. (2006). Le problème des collocations en TAL. Nouveaux cahiers de linguistiques française, 27.

VOISIN, S. (2010). Les morphèmes -*i* et -*si* en wolof STL(CLAD) (7).

WEHRLI, E. (1991). L'analyse syntaxique des langues naturelles : Problèmes et méthodes. Masson

WEHRLI, E. (2004). Un modèle multilingue d'analyse syntaxique. *In Structures et Discours. Mélanges offerts à Eddy Roulet.* Nota Bena.

WIRTH, N. (1985). ALGORITHME AND DATA STRUCTURES. HTTP://WWW.INF.ETHZ.CH/PERSONAL/WIRTH/BOOKS/ALGORITHME1/AD2012.PDF. [CONSULTE LE 28/03/2012].

## Liste des abréviations

| | |
|---|---|
| EV3S | Emphatique du verbe 3e personne du singulier sujet. |
| POSS2S | Possessif 2e personne du singulier. |
| PRES3S | Présentatif 3e personne du singulier sujet . |
| EL | Morphème de mouvement associé éloignant. |
| DEF | Déterminant défini singulier |
| DEF.P | Déterminant défini pluriel. |
| RAPP | Morphème de mouvement associé approchant |
| IMP | impératif. |
| O1S | clitique objet 1e personne du singulier. |
| O3S | clitique objet 3e personne du singulier. |
| NAR1P | narratif 1e personne pluriel sujet. |
| INACC | inaccompli. |

## Les auteurs

Mar Ndiaye est ingénieur cogniticien et informaticien linguiste formé aux technologies de la connaissance et aux technologies du langage respectivement dans les universités de Grenoble 2,3 et de Genève. Il a été assistant d'enseignement et de recherche au LATL de l'université de Genève de 2001 à 2007. Il enseigne actuellement les systèmes d'information à l'école supérieure de commerce de Dakar.

Cherif Mbodj est directeur du Centre de Linguistique Appliquée de Dakar (UCAD), Sénégal.

**Algorithme d'analyse de Fips**

**entrée**

- Soit un graphe dans lequel figurent les constituants déjà construits
- une tête de lecture qui parcourt la phrase de gauche à droite
- un agenda

**début**

Initialement, le graphe ne contient aucun élément, la tête de lecture pointe sur le premier mot de la phrase d'entrée et l'agenda est vide;

**répéter**

**Si** l'agenda est vide **alors**

Lire un mot $M$;

**pour chaque** lecture de $M$ de catégorie X **faire**

Projeter une projection maximale XP;
Inserer XP dans le graphe;
Ajouter XP à l'agenda;

**fin**

**sinon**

Extraire un constituant $C$ de l'agenda ;
Combiner $C$ avec les constituants dans son contexte gauche, à savoir pour tous les contextes gauches $G_i$ de C;
Attacher $G_i$ comme spécificateur de $C$;
/* attachement à gauche */
**pour chaque** nœud actif $A_i$ de $G_i$ **faire**

attacher $C$ comme complément de $A_i$
/* attachement à droite */

**fin**
Projeter C;
Compléter les chaînes A-barre et les chaînes clitiques ;
associées au nœud actif $A_i$ ;

**fin**

Tous les constituants résultant des opérations de combinaison, projection et complétion de chaînes sont ajoutés au graphe. De plus, ce qui résulte d'une projection ou d'un attachement à gauche sont ajoutés à l'agenda;

**jusqu'à** ce que la tête de lecture soit en fin de phrase

**fin**

algorithme 1 – coin droit