Vers la mise en place d'un lexique basé sur LMF pour la langue Wolof

Mouhamadou KHOULE [1] Mouhamad Ndiankho THIAM [1] El hadji Mamadou NGUER [1]
(1) LANI, Université Gaston Berger de Saint Louis du Sénégal, BP 234 Saint-Louis Sénégal
mouhamadoukhoule@gmail.com, thiamouhamad@gmail.com, emnguer@ugb.edu.sn

Résumé. Le Wolof est la langue la plus parlée au Sénégal mais son utilisation efficace dans l'éducation et la formation requiert le développement d'outils du TALN dont la base de travail est le lexique. Malheureusement un tel lexique n'existe pas et sa mise en place, nécessite au préalable une étude linguistique de la structuration des données de cette langue. Cependant un tel travail a été effectué pour la mise au point d'une base de données lexicale pour la langue Wolof (Cissé et al. 2007). Cette dernière se présente sous forme de fiches lexicales où des répétitions sont notées au niveau des entrées. De plus certaines informations morphologiques du lexème (formes fléchies et dérivées) n'y sont pas représentées. L'objectif de cet article est de mettre en place un lexique pour la langue Wolof en partant du travail de restructuration effectué dans (Cissé et al. 2007) mais en apportant des solutions aux problèmes cités ci-dessus.

Abstract. Wolof is the most widely spoken language in Senegal, but its effective use in education and training requires the development of NLP tools which is based on lexicon. Unfortunately such a lexicon does not exist and its implementation, requires prior linguistic study of the data structure of the language. However, such work has been done for the development of a lexical database for the Wolof language (Cissé et al. 2007). The latter is in the form of lexical records where rehearsals are noted at the inputs. In addition, some morphological information of the lexeme (inflected and derived forms) are not represented. The objective of this paper is to develop a lexicon for language Wolof starting the restructuring work done in (Cissé et al. 2007) but in providing solutions to the problems mentioned above.

Mots-clés : TALN, modèle, lexique, XML, LMF.
Keywords: **TALN**, model, lexicon, XML, LMF

1 Introduction

Le Sénégal est un pays dont 80 %[1] de la population ne comprennent pas réellement la langue officielle qu'est le français. Cela constitue un véritable handicap pour former de manière efficace la population, gage d'un développement économique réel et durable. Pour pallier à ce problème il s'avère nécessaire d'utiliser les langues nationales comme le wolof qui est compris par plus de 80 % [2]de la population.

Comparée aux langues étrangères comme le français et l'anglais, le wolof n'a pas profité des avancées du TALN dont la principale base de travail est le lexique. Notons qu'un tel lexique, qui n'est toujours pas mis en place pour la langue Wolof à l'état actuel de la recherche, requiert au préalable une étude linguistique de la structuration des données de cette langue.

Un travail de structuration de la langue Wolof a été effectué pour la mise au point de base de données multifonctionnelle pour cette langue (Cissé et al. 2007). Cette base de données lexicale est composée d'un ensemble de fiches lexicales. Néanmoins certaines informations morphologiques relatives au lexème ne sont pas disponibles sur les fiches lexicales. De plus on note beaucoup de répétitions au niveau des entrées lexicales de la base.

L'objectif de notre travail est de mettre en place un lexique pour la langue Wolof en partant du travail de restructuration effectué dans (Cissé et al. 2007). Il s'agit de structurer ces fiches lexicales suivant le standard LMF (Lexical Markcup

[1] La Francophonie dans le monde 2006-2007, éd. Nathan, Paris, mars 2007.

[2] Recensement général de la population et de l'habitat de 1988, publiés en juin 1993 par la Direction de la Prévision et de la Statistique.

Framework) qui n'est pas un format mais plutôt un méta-modèle. Ce qui nous permettra de supprimer certaines redondances mais aussi de pouvoir ajouter certaines informations morphologiques au lexème. Dans la suite du document, nous présenterons d'abord les travaux effectués dans (Cissé et al. 2007), ensuite nous parlerons du standard LMF pour enfin terminer par la structuration des fiches en suivant l'esprit LMF. L'objectif final consiste à exporter l'ensemble des fiches structurées au format LMF dans une base de données lexicale qui servira de base de travail pour la mise en œuvre d'un correcteur orthographique interactif pour la langue wolof.

2 Travaux antérieurs pour la mise en place d'une base de données lexicale pour le Wolof

Le terme Wolof désigne à la fois la langue Wolof et l'ethnie parlant le Wolof. Le wolof est la langue la plus parlée au Sénégal (par l'ethnie Wolof, environ 45 % de la population, ainsi que par les populations non-wolofs du Sénégal). Cette langue, qui est aussi parlée en Gambie et en Mauritanie, connaît une expansion culturelle fulgurante. Le wolof a longtemps été écrit avec l'alphabet arabe complété (Ajami). Cette écriture est généralement utilisée par la population formée dans les écoles coraniques (daaras), mais le wolof utilise également l'alphabet latin avec des conventions particulières pour respecter les sons particuliers de cette langue. Notons que l'alphabet latin est l'alphabet officiellement adopté par l'état. Néanmoins l'alphabet arable complété qui est aujourd'hui harmonisé est aussi reconnu par l'état. Notre travail fait référence à l'alphabet latin du Wolof.

Le Wolof, comme beaucoup de langues africaines, a connu peu d'essais d'élaboration de bases de données lexicales. Il faut saluer les quelques efforts faits jusqu'ici. A ce titre on ne retrouve que le projet de mise au point d'une base de données lexicale multifonctionnelle (Cissé et al. 2007). Il est question dans ce projet de constituer une base de données lexicale à partir de laquelle extraire à la fois un dictionnaire unilingue wolof et un dictionnaire bilingue wolof/français. Il se fixe parmi ses objectifs de produire des sorties XML et de concevoir des modèles XSL pour l'interrogation.

Le schéma descriptif des entrées repose sur une hiérarchisation en trois niveaux des données. Cette hiérarchisation permettra, entre autres, d'utiliser le dictionnaire avec un degré de granularité différent selon les besoins des usagers. Au premier niveau d'information, qui correspond au champ de la lexie, sont associées les informations hiérarchisées sur deux autres niveaux comme suit :

- champs secondaires : information qualifiant directement le champ primaire « lexème », telles les données se rapportant à la « catégorie grammaticale » ou aux « synonymes ».

- champs tertiaires : information qualifiant une donnée secondaire. Par exemple, le champ « classe nominale » est un champ subordonné du champ « catégorie grammaticale ».

La figure 1 présente une illustration d'une entrée ainsi que les champs qui lui sont associés (Cissé et al. 2007). L'image est obtenue à partir de l'outil Toolbox que les concepteurs ont utilisé pour la conception de la base de données. Ce qui explique la présence des champs d'administration (statut de la fiche, commentaires, auteur du statut de la fiche).

\lex Lexème wolof	askan
\uttW Transcription phonétique	eskeN
\fsLW Fichier son du lexème wolof	C:\Dictionnaire_Wolof\askan_population.wav
\catW Catégorie grammaticale du lexème wolof	turu bokkaale
\clasW Classe nominale du lexème wolof	w-
\srcLW Source du lexème wolof	
\defW Définition du lexème wolof	Mbooleem ñi bokk dëkkandoo
\srcDW Source de la définition du lexème wolof	
\attW Contexte d'attestation du lexème wolof	Texte juridique
\srcAW Source du contexte d'attestation du lexème wolof	Déclaration universelle des droits de l'homme (http://www.unhchr.ch/udhr/lang/wol.htm)
\nusW Note d'usage du lexème wolof	
\varW Variante du lexème wolof	
\synW Synonyme du lexème wolof	
\homW Homonyme du lexème wolof	askan
\homW Homonyme du lexème wolof	askan
\exDerW Expression dérivée du lexème wolof	
\lexSrcW Lexème source de l'expression dérivée	
\CA Corpus associé	CC
\tradFlex Traduction francaise du lexème wolof	Population
\catF Catégorie grammaticale de la traduction francaise	nom
\phrW Phrase d'illustration du lexème wolof	Njaboot nekk na meññeef gu am solo ci askan wi.
\fsPhrW Fichier son de la phrase d'illustration	C:\Dictionnaire_Wolof\askan_population_phr.wav
\tradPhrW Traduction francaise de la phrase d'illustration	La progéniture constitue une ressource importante pour la population.
\stat Statut de la fiche	ok
\cmt Commentaire	
\autStat Auteur du statut de la fiche	AMD
\dat Date de dernière modification de la fiche	10/Apr/2008

Figure 1: exemple de fiche lexicale complète obtenue avec l'outil Toolbox

Bien que l'envergure de ce projet soit grande, au niveau du modèle on se rend compte que l'on a affaire à des concepts assez simples. En effet la structuration est celle d'une fiche. On a une liste de fiches avec tous les champs nécessaires et des renvois possibles entre fiches (synonymie, homonymie). Les concepteurs ont pris un certain nombre de dispositions vis-à-vis des spécificités de la langue Wolof. Par exemple au niveau des entrées on note beaucoup de répétitions, chose qu'ils justifient par les besoins de différentiation par les termes suivants (Cissé et al. 2007): "S'agissant d'une base de données informatisée, nous avons volontairement privilégié une « structuration monosémique » afin de répondre adéquatement aux exigences de l'ingénierie linguistique. Dans la pratique, cela signifie qu'une lexie wolof polysémique (à laquelle correspond nécessairement plus d'un équivalent en français) fera l'objet de plusieurs entrées''. De plus certaines informations morphologiques du lexème telles que les formes dérivées et fléchies ne sont pas disponibles dans la fiche. Dans la partie suivante, nous allons restructurer les fiches en suivant le standard LMF tout en y ajoutant certaines informations morphologiques relatives au lexème. Ceci va nous permettre aussi de supprimer certaines redondances au niveau des entrées.

3 Vers une élaboration du lexique basé sur LMF

3.1 Choix et présentation générale de LMF

3.2 Choix de LMF

Concernant les standards, nous avons porté notre choix sur LMF (Lexical Markup Framework) devenu norme ISO numéro 24613 :2008 en novembre 2008 (Enguehard et al. 2011) pour plusieurs raisons. Tout d'abord les objectifs de LMF sont de fournir un modèle commun pour la création et l'utilisation de ressources lexicales, mais aussi de permettre l'interopérabilité entre ces ressources (Francopoulo et al. 2006). Elle permet la spécification de ressources linguistiques monolingues et multilingues destinées à l'usage éditorial et du TALN. Les langues couvertes par LMF ne se limitent pas aux langues européennes mais à toutes les langues naturelles. De plus elle assure une modélisation extensible et modulaire couvrant tous les niveaux de description linguistique (morphologique, syntaxiques, sémantique, etc.).

3.3 Présentation générale de LMF

LMF est une initiative au sein de l'ISO en faveur de la normalisation de la représentation des ressources lexicales. A partir des expériences acquises au cours des études antérieures (Genelex, EAGLES, ISLE, Multext, TEI), l'idée est de proposer un modèle de données modulaire, indépendant vis-à-vis d'une théorie lexicographique particulière et permettant de s'abstraire de la représentation concrète (SGML/XML, DTD propriétaire ou TEI, base de données relationnelle, etc.).
LMF propose un méta-modèle constitué d'un noyau obligatoire autour duquel gravitent des extensions (morphologique, syntaxique, sémantique et MRD) (Francopoulo et al. 2006). Le noyau de LMF est présenté par la figure 2. L'objet «Lexical Entry » contient un ou plusieurs objets « Form » et un ou plusieurs objets « Sense».

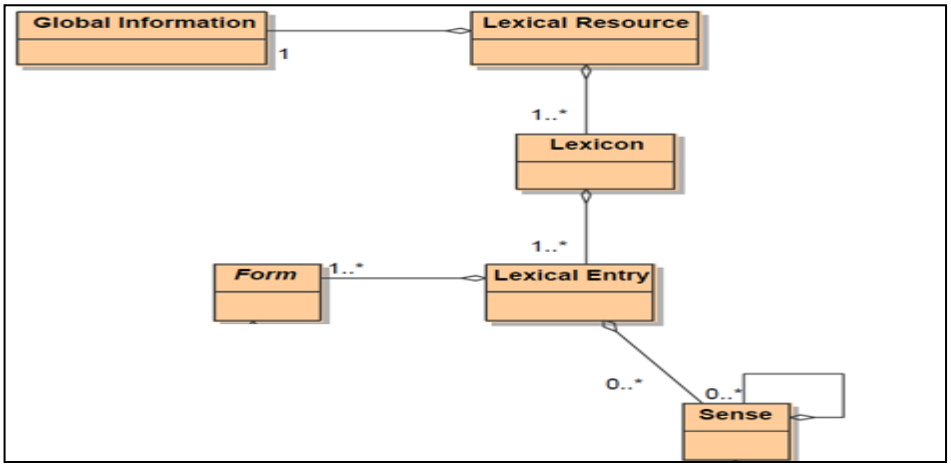

Figure 2 : Noyau du méta-modèle LMF

Nous allons maintenant structurer nos fiches en suivant ce méta-modèle.

3.4 Structuration des fiches en suivant l'esprit LMF

Les fiches produites dans les travaux dans (Cissé et al.2007) sont disponibles au format XML. Nous allons maintenant les structurer en suivant l'esprit LMF. La figure 3 présente une fiche lexicale après l'ajout des balises de structuration. La figure 4 présente la fiche au format LMF. La balise « fiche » correspondant à l'objet « Lexical Entry », la balise « bloc-vedette » correspond à l'objet « form» et la balise «bloc-sémantique » correspond à l'objet «Sense ». Nous allons juste prendre les informations dont nous avons besoin au niveau de la fiche lexicale. Nous ajouterons ensuite certaines balises de restructuration pour que nos fiches prennent en compte les formes fléchies et les formes dérivées.

```
<fiche>
        <bloc-vedette><lex>askan</lex><uttW>ɐskɐn</uttW><bloc-vedette>
        <bloc-grammatical>
                <catW> turu bokkaale</cat>
                <clasW> w- </clasW>
                <bloc-sémantique>
                        <defW>Mbooleem ñi bokk dëkkandoo</defW>
                        <attW>Texte juridique</attW>
                        <srcAW>Déclaration universelle des droits de l'homme
(http://www.unhchr.ch/udhr/lang/wol.htm)</srcAW>
                        <phrW>Njaboot nekk na meññeef gu am solo ci askan wi.</phrW>
                        <tradFlex>population</tradFlex>
                        <catF>nom</catF>
                        <tradPhrW>La progéniture constitue une ressource importante pour la
population<tradPhrW>
                        <synW></synW>
                        <homW>askan</homW>
                        <homW>askan</homW>
                </bloc-sémantique>
        </bloc-grammatical>
</fiche>
```

Figure 3: fiche après structuration

Après avoir structuré nos fiches en bloc, nous allons les restructurer en suivant le standard LMF. La balise /WordForm permet de prendre en charge les formes fléchies et la balise /RelatedForm les formes dérivées.

```
<LexicalEntry id="1">

                    <feat att="partOfSpeech" val="noun"/>

                     <feat att="affixClass" val="w-"/>

 <Lemma><fea att="writtenForm" val="askan"/><fea att="phoneticForm" val="ɐskɐn"/></Lemma>

<WordForm></WordForm>

<Sense>

        <Definition> <feat att="writtenForm" val="Mbooleem ñi bokk dëkkandoo"/> <feat att="source"
        val=""/></Definition>

        <Contecxt><feat att="text" val=Njaboot nekk na meññeef gu am solo ci askan wi."/></Context>

        <Context><feat att="language" val="fra"/><feat att="text" val="La progéniture constitue une
ressource importante pour la population"></Context>

        <SubjectField><feat att="writtenForm" val="Texte juridique"/><feat att="source"
val="Déclaration universelle des droits de l'homme (http://www.unhchr.ch/udhr/lang/wol.htm)"/>
</SubjectField>

        <Equivalent> <feat att="language" val="fra"/> < feat att="partOfSpeech" val="noun"/><feat
att="writtenForm" val="population"/></Equivalent>

        <SenseRelation target=""><feat att="label" val="synonym"/></SenseRelation>

        <SensRelation target="askan"><feat att="label" val="homonym"></SenseRelation>

        <SensRelation target="askan"><feat att="label" val="homonym"></SenseRelation>

 </Sense>

 <RelatedForm> </RelatedForm>

</LexicalEntry>
```

Figure 4: Fiche au format LMF

4 Conclusion

L'objectif des travaux présentés dans cet article est de mettre en place un lexique basé sur LMF (Lexical Markup Framework) pour la langue Wolof parlée par près de 80% la population sénégalaise. Nous avons pour cela fait un état de l'art des bases de données lexicales en Wolof. Ce qui nous a permis de constater que les travaux allant dans ce sens restent uniquement ceux dans (Cissé et al. 2007) dont le but principal est l'étude de la structuration de la langue Wolof et la mise au point de base de données multifonctionnelle pour cette langue.

Cette base de données lexicale, composée d'un ensemble de fiches lexicales, présente néanmoins quelques inconvénients : certaines informations morphologiques relatives au lexème ne sont pas disponibles sur les fiches et on note aussi beaucoup de répétitions au niveau des entrées lexicales de la base. Cependant elle permet de produire des sorties XML des fiches lexicales.

Pour obtenir un tel lexique, nous avons d'abord restructuré ces fiches lexicales en différents blocs pour ensuite proposer une méthode de conversion de ses fiches lexicales en suivant le standard LMF, tout en y ajoutant certaines balises pour la prise en charge des formes fléchies et dérivées relatives au lexème.

Ce travail de mise en place d'un lexique pour la Wolof est très bénéfique dans la mesure où il constitue une base de travail nécessaire pour développer un correcteur interactif et un traducteur automatique pour cette langue.

Dans nos futurs travaux, nous comptons automatiser la structuration des fiches selon LMF en utilisant une feuille de style XSLT, pour mettre en place une base de données lexicale normalisée LMF pour la langue Wolof, concevoir un outil d'intégration des différentes fiches lexicales structurées suivant l'esprit LMF et un outil d'enrichissement et d'interrogation de la base de données normalisée.

5 Bibliographie

Cisse M.T., Thiaw N. F. (2007) Le projet de dictionnaire unilingue wolof et bilingue wolof-français : une base de données lexicale. Actes des Journées ASR 2007, Tunis.

Cisse M.T., Diagne A.M., Campenhoudt M.V., Muraille P. (2007) Mise au point d'une base de données lexicale multifonctionnelle : le dictionnaire unilingue wolof et bilingue wolof-français. Actes des Journées LC 2007, Lorient.

Enguehard C., Mangeot M. (2011) Informatisation de dictionnaires langues africaines-français. Actes des journées LTT 2011, Villetaneuse.

Francopoulo G., George M., Calzolari N, Monachini M., Bel N., Pet M., Soria C. (2006) Lexical Markup Framework (LMF). LEREC, Genoa.

Baccar F., Khemakhem A., Gargouri B., Haddar K., Hamadou Abdelmajid B. (2008) Modélisation normalisée LMF des dictionnaires électroniques éditoriaux de l'arabe. TALN 2008, Avignon,France