Dictionnaires wolof en ligne : État de l'art et perspectives

El Hadji Nguer, Mouhamadou Khoulé, Ousmane Thiaré, Mame Thierno Cissé, Mathieu Mangeot

Dictionnaires wolof en ligne :

État de l'art et perspectives

El hadji Mamadou Nguer, Mouhamadou Khoulé, Ousmane Thiaré* -- Mame Thierno CISSE**-- Mathieu MANGEOT***

*LANI
Université Gaston Berger (UGB), SENEGAL
emnguer@ugb.edu.sn, mouhamadoukhoule@gmail.com, ousmane.thiare@ugb.edu.sn

**Université Chaikh Anta Diop (UCAD), SENEGAL
thiernoc@gmail.com

***Université de Savoie, FRANCE
mathieu.mangeot@imag.fr

RÉSUMÉ. Dans beaucoup de pays francophones d'Afrique de l'Ouest, les populations dans leur majorité utilise les langues locales. C'est le cas du Sénégal où plus de 80% de la population utilise dans ses activités au quotidien, le wolof, langue endogène et véhiculaire. Mais cette langue est malheureusement peu dotée d'outils modernes comme des dictionnaires normalisés, nécessaires à leur introduction efficace dans le système éducatif pour la prise en charge des besoins en formation des populations. Le présent travail fait un état de l'art sur les dictionnaires en ligne existants sur le wolof en prélude à la mise en place d'un dictionnaire en ligne normalisé en construction collaborative sur le web répondant au standard LMF (Lexical Markup Framework) [3].

ABSTRACT. In many francophone countries in West Africa, people in their majority use local languages. This is the case of Senegal, where more than 80% of the population uses in its daily activities, Wolof, endogenous and vehicular language. But this language is unfortunately not equipped with modern tools such as standardized dictionaries, for their effective introduction into the education system for the care of people's training needs. This work is a state of the art on existing online dictionaries Wolof as a prelude to the establishment of a standardized online collaborative dictionary construction on the web addressing the standard LMF (Lexical Markup Framework) [3].

MOTS-CLÉS : Langue, wolof, traitement automatique des langues, dictionnaire informatique, LMF.

KEYWORDS: Language, wolof, Natural language processing, electronic dictionary, LMF.

1. Introduction

AU Sénégal, vingt-cinq (25)[1] langues endogènes cohabitent avec le français, l'arabe et d'autres langues étrangères. Il est avéré que le français et le wolof dominent largement dans les transactions langagières sur l'étendue du territoire national. Le français, qui est la langue officielle, n'est compris réellement ou occasionnellement que par près de 30% sur une population[1] de 14 133 280. Cela veut dire que 9 856 280 sénégalais ne comprennent le français. Ce qui constitue un handicap important si l'on veut informer et former la population de manière efficace, gage d'un développement socio-économique réel et durable.

En réalité, comparée aux langues étrangères comme le français et l'anglais, nos langues nationales à l'instar du wolof, qui est la langue la plus parlée au Sénégal, n'ont pas pu bénéficier des avancées du TALN (Traitement automatique des langues naturelles). Jusqu'à présent le wolof n'est pas encore doté d'outils et de ressources tels que des dictionnaires normalisés, des correcteurs orthographiques, sans parler de traducteurs automatiques. Ces outils sont nécessaires pour accompagner l'introduction généralisée de nos langues nationales dans l'éducation et la formation au Sénégal. En faisant le point sur l'état de l'art des dictionnaires en ligne sur la langue wolof, le présent travail se veut une contribution au travail préalable d'utilisation optimale des langues nationales pour le développement socio-économique du pays.

La suite du document est composée comme suit :

- la première partie sera consacrée à une brève présentation de la langue wolof ;
- un rappel des notions de dictionnaire sera fait dans la deuxième partie avec une présentation générale de LMF ;
- l'état de l'art des dictionnaires en ligne sur cette langue sera abordé dans la troisième partie ;
- enfin dans la quatrième partie, nous proposerons une solution en cours d'élaboration. Cette solution consiste à mettre en place un dictionnaire en construction collaborative sur le web, basé sur la norme LMF à partir de la base de données multifonctionnelle pour la langue Wolof (Cissé et al. 2007) [2]. Il sera ensuite question de dégager les perspectives.

[1] Selon la Direction de l'Alphabétisation et des Langues Nationales au Sénégal.

2. Présentation de la langue wolof.

Le terme wolof désigne à la fois la langue wolof et l'ethnie qui parle cette langue. Majoritairement parlé au Sénégal (par l'ethnie Wolof, environ 45 % de la population, ainsi que par les populations non-wolophones du Sénégal), cette langue a le statut de langue nationale au Sénégal, en Gambie et en Mauritanie.

Le wolof est officiellement écrit avec l'alphabet latin avec des conventions particulières pour respecter les sons particuliers de la langue. Mais elle est aussi écrite avec l'alphabet arabe complété (Ajami).

De nos jours, la vitalité du wolof s'accentue, notamment grâce à l'urbanisation; parler le wolof lorsqu'on vit dans des villes comme Dakar, Louga, Thiès, Saint-Louis ou Kaolack est indispensable. Par ailleurs,

- le wolof s'impose de plus en plus dans les débats télévisés, les émissions radio et les panneaux publicitaires;

- le wolof fait partie des langues nationales ayant le plus fait l'objet d'études et de recherches;

- le wolof est présent sur Wikipedia, sur les outils de Windows et sur ceux de Google;

- la constitution du Sénégal, le code des marchés, le coran et la bible sont entièrement traduits en wolof;

- le wolof fait partie des langues transfrontalières véhiculaires choisies par l'Académie Africaine des Langues (Acalan).

Malgré tous ces atouts, le wolof n'est toujours pas doté de dictionnaire en ligne normalisé.

3. Notions de dictionnaire

Un dictionnaire est d'une importance sociale considérable. En entreprenant la rédaction d'un dictionnaire on peut même affirmer l'existence d'une société, d'une culture [8].

3.1. Qu'est-ce qu'un dictionnaire ?

D'une manière générale, un dictionnaire est un précieux et excellent outil d'apprentissage d'une langue. Il s'agit d'un ouvrage de référence qui répertorie les mots d'une langue dans un ordre convenu (alphabétique en général) pour leur associer par exemple une définition, une étymologie, une traduction etc. Il est un répertoire du lexique de la langue. Il est composé d'un ensemble de volumes. Chaque volume est composé d'un ensemble d'articles. La liste ordonnée de ces articles constitue la nomenclature du dictionnaire. L'ordre utilisé est généralement l'ordre alphabétique des mots-vedettes de la langue. Un article est composé d'un mot-vedette (appelée aussi entrée ou terme) et d'un corps.

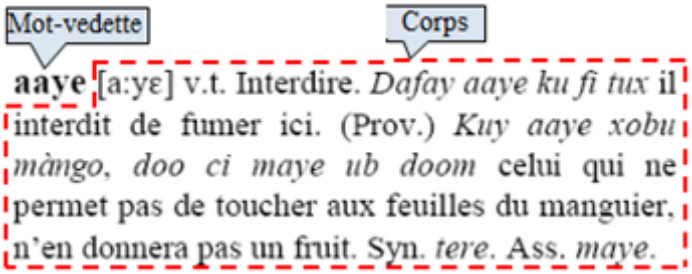

Figure 1 Un exemple d'article du dictionnaire Wolof-Français de Jean Léopold Diouf

3.2. Les types de dictionnaire

Il existe deux types de dictionnaires : les dictionnaires terminologiques et les dictionnaires généraux.

- Un dictionnaire terminologique rassemble généralement les entrées d'un domaine précis de la langue.

- Un dictionnaire général rassemble au contraire tous les termes de la langue sans se spécialiser dans un domaine particulier. Il contient généralement des informations assez riches et variées.

3.3. La macrostructure des dictionnaires

L'organisation des volumes du dictionnaire constitue la macrostructure du dictionnaire.

La macrostructure la plus simple est celle qui est faite en un seul volume. Dans les dictionnaires composés d'un seul volume, les mots-vedettes appartiennent à la même langue. Ces dictionnaires sont principalement des dictionnaires monolingues ou bilingues monodirectionnels (langue A vers langue B). Il existe également des dictionnaires multilingues indexés selon une seule langue. Ce sont les dictionnaires multicibles ou furcoides [6].

Une macrostructure fréquemment utilisée est celle du dictionnaire bilingue en deux volumes, l'un trié selon les mots-vedettes d'une langue source et traduisant ces mots-vedettes dans une autre langue cible et l'autre volume symétrique. Ils sont appelés dictionnaires bilingues bidirectionnels.

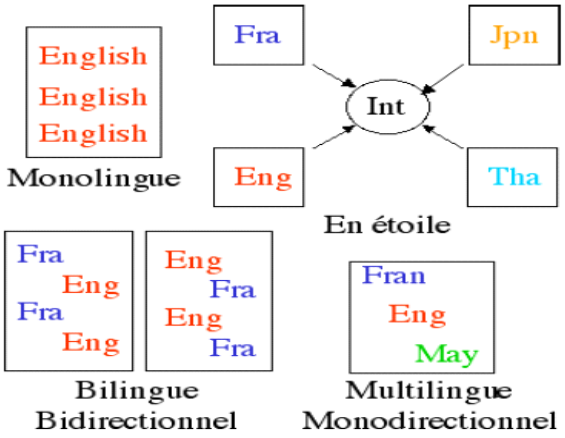

Figure 2: Exemples de macrostructures

Une autre macrostructure plus complexe destinée aux bases de données multilingues consiste à organiser en étoile autour d'un dictionnaire central de concepts ou d'acceptions, des dictionnaires monolingues contenant dans chaque langue de la base

les traductions des concepts ou acceptions du dictionnaire central [6]. Le dictionnaire central joue le rôle de pivot de la base. La Figure 2 tirée de la thèse de Mathieu Mangeot [6] représente les principaux exemples de macrostructures.

3.4. La microstructure des dictionnaires

La structure logique de l'article constitue la microstructure du dictionnaire. Nous pouvons la considérer comme une structure composée d'objets linguistiques. Parmi ces objets linguistiques nous pouvons citer le mot-vedette, la prononciation, la catégorie grammaticale du mot vedette (nom, verbe, adjectif, etc.), la classe nominale, la définition, la traduction, le sens, les dérivés, l'homonyme, le synonyme, l'antonyme, etc.

3.5. Les normes de présentation de dictionnaires.

Nous ne pouvons pas aborder les notions de dictionnaire sans mentionner les normes de structuration de dictionnaires qui est régie par des standards à l'instar de LMF [3] ou TEI [9]. Concernant ces standards, nous avons porté notre choix sur LMF (Lexical Marckup Framework) devenu norme iso numéro 24613 :2008 en novembre 2008[5] 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) [3]. Elle permet la spécification de ressources linguistiques monolingues et multilingues destinées à l'usage éditorial et du traitement automatique de la langue naturelle (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.).

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) [3]. Le noyau de LMF est présenté à la

Figure 3. L'objet «Lexical Entry » contient un ou plusieurs objets « Form » et un ou plusieurs objets « Sense».

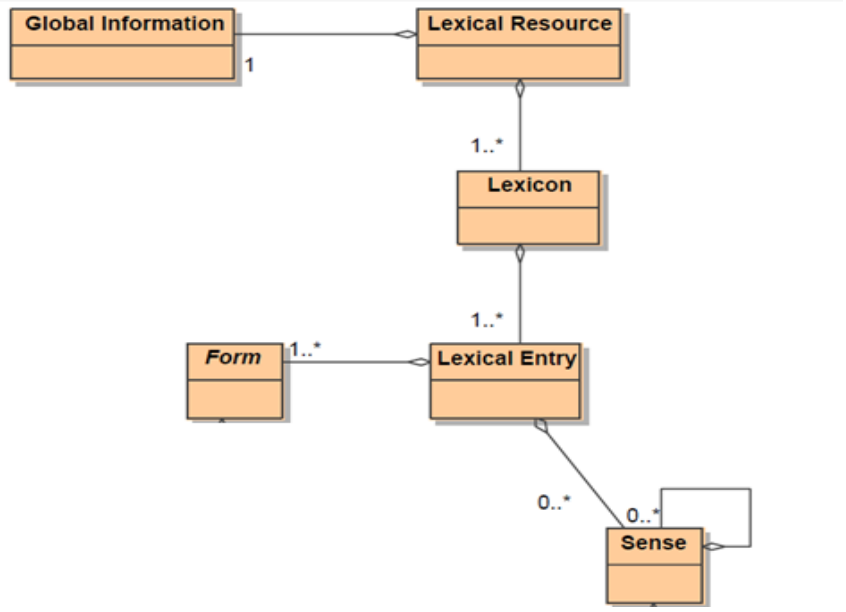

Figure 3 Noyau du méta-modèle LMF

4. Etat de l'art des dictionnaires en ligne du wolof

Il s'agit, dans cette partie, de présenter la liste des dictionnaires informatiques en ligne existants sur le wolof en mettant en exergue leurs caractéristiques (macrostructure, microstructure, fonctions de recherche et de mise à jour, etc.), leurs atouts ainsi que leur limite. Notons d'emblée qu'ils sont tous de type général.

4.1. Le projet de dictionnaire unilingue wolof et bilingue wolof-français de Cissé & al [1].

Ce projet financé par l'Agence Universitaire de la Francophonie (AUF), a réuni le département de linguistique de l'Université Cheikh Anta Diop de Dakar (Sénégal),

le Centre de recherche Termisti de l'Institut supérieur de traducteurs et interprètes, Haute École de Bruxelles (Belgique) et l'Institut für Linguistik/phonetik de l'université de Cologne (Allemagne).

Il est question dans ce projet de constituer une base de données lexicale à partir de laquelle il est possible d'extraire à la fois un dictionnaire unilingue wolof et un dictionnaire bilingue wolof/français.

Il se fixe parmi ses objectifs :

- De produire une sortie au format XML pour la réutilisation dans des outils d'ingénierie linguistique, ainsi que des modèles XSL permettant à quiconque de consulter le dictionnaire en ligne ou hors ligne.

- D'étudier la faisabilité de la production d'un correcteur orthographique intégré (MySpell / OpenOffice) basé sur le dictionnaire.

L'encodage des données lexicographiques s'est effectué à l'aide du gratuiciel Toolbox[2] (version 1.5) de SIL international Le modèle de données retenu privilégie une approche monosémique de manière à garantir au mieux l'établissement des équivalences et à demeurer compatible avec les exigences de l'ingénierie linguistique.

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 4 présente une illustration d'une entrée ainsi que les champs qui lui sont associés. L'image est obtenue à partir de l'outil Toolbox.

[2] http://www.sil.org/computing/toolbox.

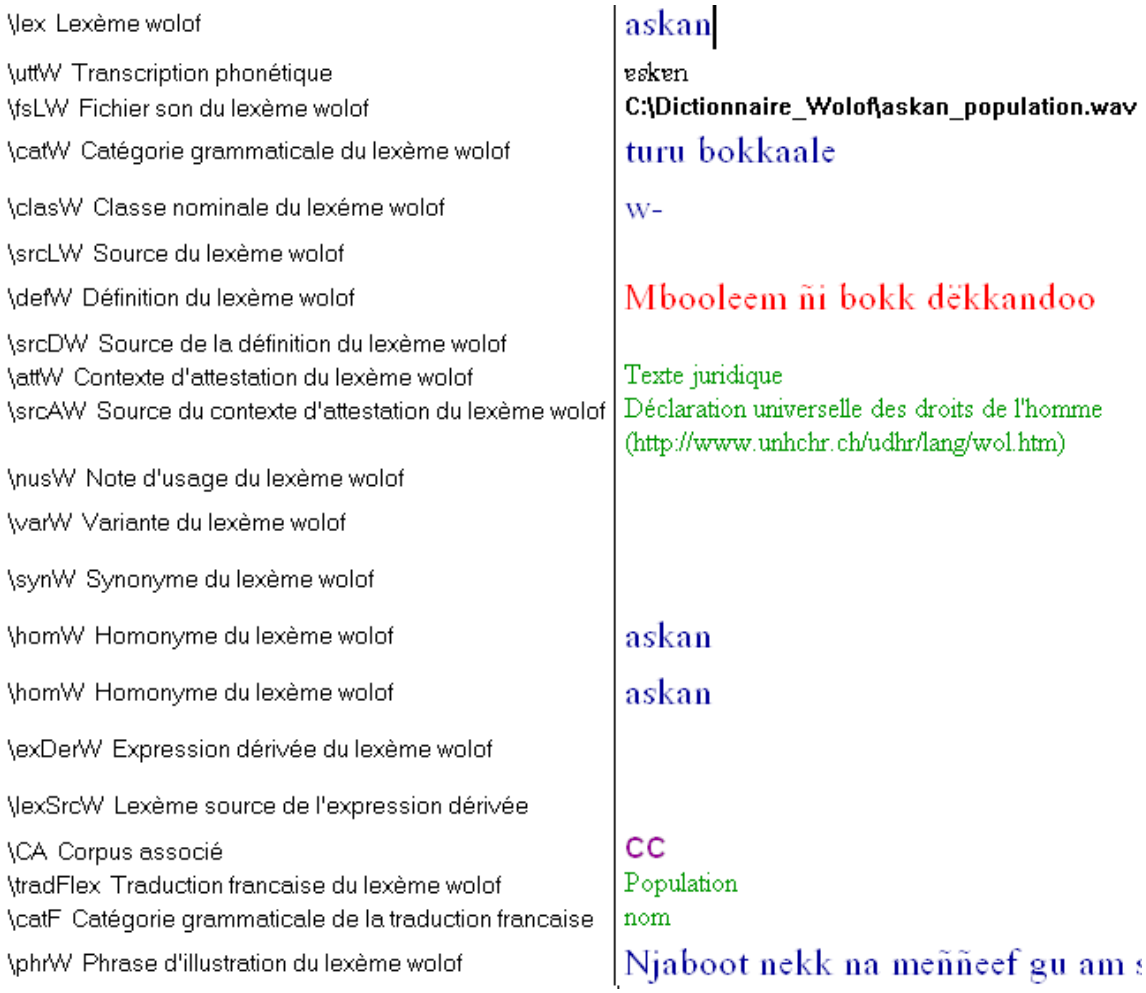

Figure 4: Exemple de fiche lexicale obtenu avec l'outil Toolbox

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.

Au-delà des redondances on peut remarquer le manque d'utilisation de format de représentation normalisée de dictionnaire à l'instar de LMF [3] ou TEI [9]. L'échantillon du dictionnaire est disponible à l'adresse http://flsh-dico-wolof.ucad.sn/xml/A-Z-wo-sort-01.xml, qui malheureusement ne présente pas de fonction de recherche et de mise à jour du dictionnaire.

Cependant ce projet a le mérite d'avoir permis d'effectuer une bonne structuration du wolof et de germer une base de données lexicale de plus 8 167 entrées, ayant une microstructure proposée et validée par des experts du domaine.

4.2. Le dictionnaire Freelang wolof-français

Le dictionnaire Freelang est un projet contributif, auquel tous les utilisateurs peuvent participer. L'objectif est de mettre à disposition sur Internet, gratuitement, un maximum de lexiques bilingues, eux-mêmes composés d'un maximum de mots traduits de la manière la plus exacte possible.
Freelang se propose de tendre vers cet objectif :

- en mettant à la disposition des internautes un programme gratuit et des listes de mots facilement modifiables par le biais de ce programme (grâce à ses fonctions d'ajout, de modification ou de suppression de traductions) ;
- en aidant les utilisateurs à créer de nouvelles listes de mots pour le dictionnaire Freelang ;
- en intégrant aux listes de mots existantes les mises à jour envoyées par les utilisateurs ;
- en convertissant au format du dictionnaire Freelang les lexiques déjà réalisés par des utilisateurs sous d'autres formats.

Le dictionnaire Freelang propose des dictionnaires de plusieurs langues. Parmi lesquelles, il existe le dictionnaire bilingue bidirectionnel wolof-français. La

Figure 5 constitue une illustration du dictionnaire.

Figure 5 : Traduction en français du mot wolof fàtte.

Même si le projet est très prometteur dans le domaine de la coopération sur internet, on regrette le manque de qualité des données vu qu'on a affaire à un travail de bénévolat (traducteurs) sans la validation par des lexicographes. En outre la microstructure du dictionnaire est très pauvre et de plus il n'est pas possible de télécharger la source du dictionnaire dans un seul fichier.

4.3. Autres dictionnaires informatiques du wolof

En plus des dictionnaires énumérés ci-dessus, il existe d'autres dictionnaires de la langue wolof qui ne permettent pas non plus le téléchargement des sources. Certains sont des dictionnaires papiers numérisés à l'instar du dictionnaire wolof-français et français-wolof de Jean Léopold Diouf et le dictionnaire français-wolof et français-bambara, par Jean Dard (1825). L'inconvénient avec ces dictionnaires papiers numérisés c'est quand, à partir de la version numérisée, si on cherche un mot-vedette dans le dictionnaire toutes les pages contenant ce mot s'affichent.

Les autres dictionnaires consultables en ligne se présentent sous la forme d'une liste de mots avec leurs définitions et éventuellement leurs traductions vers une autre langue (le français en général). Parmi ces dictionnaires on peut citer : le dictionnaire AfroWeb, le petit wolof, Wiktionary en wolof, …

5. Perspectives

Après avoir fait le tour des dictionnaires en ligne du wolof, il s'avère que des efforts importants ont été réalisés notamment avec le projet de Cissé & al. En effet, il a permis d'exploiter plusieurs dictionnaires papiers dont celui de Jean Léopold Diouf, d'effectuer une bonne structuration du wolof et d'avoir germé une base de données lexicale de plus de 8167 entrées, de proposer une microstructure validée par des experts du domaine. Ce qui est loin d'être le cas pour les autres dictionnaires tels que Wiktionary, Afro Web, Petit wolof, etc. qui restent de belles initiatives. Pour la mise en place d'un dictionnaire collaboratif en ligne basé sur LMF pour le wolof, nous avons porté notre choix sur le dictionnaire de Cissé & al. comme base de travail.

Ainsi nous avons récupérer les données du projet de dictionnaire au format XML [1]. Comme on peut le voir à la Figure 6 les noms de balises ne sont pas dans un langage compréhensible.

```
<lexGroup>
<lex>fii ak</lex>
<uttW>fi:ek'</uttW>
<varW>feek</varW>
<varW>fileek</varW>
<aut>NFT</aut>
<dat>03/Sep/2007</dat>
</lexGroup>
```

Figure 6: Article au format XML du dictionnaire Cissé & al

Ainsi nous allons successivement :

1. remplacer les noms de balises en des noms compréhensibles en français et en wolof. Ce qui nous permettra de créer une terminologie de balises en wolof,
2. structurer les articles vers un format plus standardisé (voir Figure 7). Pour cela nous allons utiliser les expressions régulières présentes dans certains langages comme Perl, Java, etc. Ensuite nous allons regrouper les informations telles que les

catégories lexicales autour d'une balise ***bloc_forme***. Les informations relatives à la sémantique autour d'une balise ***Sens*** et celles relatives aux méta-informations dans une balise **bloc_métainformation**. Et enfin les différentes dérivées seront regroupées dans une balise ***bloc_dérivés***,

```
<article id="barigo1">
<bloc_forme>
<mot_vedette>barigo</mot_vedette>
</bloc_forme>
<bloc_métainformation>
<auteur>MTC</auteur>
<date_dernière_modification>02/Sep/2007</date_dernière_modification
</bloc_métainformation>
</article>
```

Figure 7: Article dans un format plus structuré

3. affecter un identifiant unique à chaque article et à chaque sens,
4. faire le tri des articles en utilisant un script PERL,
5. Avec un script perl fusionner aussi les articles de même catégorie grammaticale, ce qui nous permettra de supprimer les redondances,
6. utiliser un fichier XSLT pour convertir automatiquement le dictionnaire en LMF.
7. mettre en ligne le dictionnaire dans la plateforme jibiki [7]. À partir de ce moment tous les acteurs (lexicologues, lexicographes, etc) pourront contribuer en ligne et même exporter les données à d'autres fins. Jibiki est une plate-forme générique en ligne pour manipuler des ressources lexicales avec gestion d'utilisateurs et groupes, consultation de ressources hétérogènes et édition générique d'articles de dictionnaires. La plate-forme est programmée entièrement en Java, basée sur l'environnement "Enhydra". Toutes les données sont stockées au format XML dans une base de données (Postgres).
8. Produire à partir du dictionnaire un lexique de formes fléchies en prélude de la mise en place d'un correcteur orthographique.

Notons que les points 1,2, 3, 4 & 5 sont déjà réalisés et que les point 6 & 7 sont est en cours de réalisation.

6. Conclusion

Il est certainement temps de doter le wolof (langue véhiculaire du Sénégal parlée par plus de 80% de la population) d'un dictionnaire en ligne, à la hauteur de ce que représente la langue au Sénégal. L'objectif de ce présent travail est de faire l'état de l'art des dictionnaires en ligne existants pour la langue afin de justifier le projet de mise en place d'un dictionnaire collaboratif en ligne répondant aux normes LMF[3] pour le wolof.

Ce projet, en cours d'élaboration, permettra à moyen terme de transformer la base de données multifonctionnelle de [2] en un format respectant la norme LMF, qui sera ensuite utilisé pour mettre en place un dictionnaire collaboratif en ligne. Il donnera aux acteurs (linguistes, lexicologues, lexicographes, chercheurs, etc.) la possibilité d'ajouter de nouvelles entrées, de mettre à jour les entrées et d'exporter la base de données.

Ce dictionnaire en LMF sera par la suite utilisé pour produire un lexique de formes fléchies qui sera utilisé à des fins de correcteur orthographique, pour la traduction automatique mais aussi pour faire du Linked Data [10], la publication des données de manière structurée sur le Web.

Dans un autre registre, ce dictionnaire sea doté d'un module de transittération automatique entre les deux écritures du wolof (écriture avec des caractères latins et écriture avec des caractères Ajami). En fait, les termes du dictionnaire stockés en caractères latins, seront présentés :

- en caractères latins pour les populations utilisant l'alphabet wolof latin
- et en caractères Ajami pour les populations utilisant l'alphabet Ajami

Ce module permettra de bénéficier de l'expertise collaborative des populations sans distinction de graphie. Il s'agira, dans la graphie de son choix, de proposer des entrées, de poster des sujets dans des forums de discussion et d'exporter la base de données à d'autres fins.

7. Bibliographie

[1] 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.

[2] 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.

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

[4] Khoulé M., Nguer E.M., Thiam M. D. (2014). Vers la mise en place d'un lexique basé sur LMF pour la langue wolof. TALN-RECITAL Workshop TALAf 2014 : Traitement Automatique des Langues Africaines (TALAf 2014: African Language Processing). Marseille, France (Association pour le Traitement Automatique des Langues) P 172—177.

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

[6] Mangeot M. Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue. Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, Grenoble, France, 2001.

[7] Mangeot M., Sérasset G., Lafourcade M. (2003) Construction collaborative de données lexicales multilingues : le projet Papillon. Revue TAL, Vol. 44:2/2003, pp. 151-176.

[8] Polguère A*,'la lexicographie", "dans Notions de base en lexicologie '', Montréal : Québec,2002,pp. 175-183.*

[9] TEI P4. Guidelines for Electronic Text Encoding and Interchange. 2004.www.tei-c.org/release/doc/tei-p4-doc/html/

[10] Heath T. and Bizer C., Linked Data: Evolving the Web into a Global Data Space, 2011, Morgan & Claypool, http://linkeddatabook.com/editions/1.0/