Manipulation de dictionnaires d'origines diverses pour des langues peu dotées : la méthodologie iBaatukaay

Mouhamadou Khoulé, Mathieu Mangeot, Mamadou Nguer

Manipulation de dictionnaires d'origines diverses pour des langues peu dotées : la méthodologie iBaatukaay

Mouhamadou KHOULE[1], Mathieu Mangeot[2], El hadji Mamadou NGUER[1]

(1) LANI, Université Gaston Berger, BP 234 Saint Louis, Sénégal
(2) LIG, Université de Grenoble Alpes, 38400 Saint Martin D'HERES, France.

mouhamadoukhoule@gmail.com,mathieu.mangeot@imag.fr,emnguer@ugb.edu.sn,

RÉSUMÉ

En général les langues africaines sont des langues peu dotées. La plupart des ressources existantes n'existent qu'au format papier. Il y a une rareté d'outils informatiques pour ces langues. C'est pour apporter des solutions à ces problèmes que le projet iBaatukaay est lancé. Son objectif est de mettre en place une base lexicale multilingue contributive sur le Web pour les langues africaines notamment sénégalaises (wolof, pulaar, bambara, etc.). Le projet doit être une base pour la constitution de correcteurs orthographiques, de traducteurs automatiques et autres dictionnaires électroniques. iBaatukaay se veut utile et ouvert à la collaboration de toutes les personnes ayant un intérêt pour les langues concernées et les données produites seront téléchargeables gratuitement sous licence Creative Commons.

ABSTRACT

Generally, African languages are less-resourced languages. Most of the existing resources exist only in printed version. There is a scarcity of IT tools for these languages. iBaatukaay projet is launched to provide some solutions to these problems. The aim of the iBaatukaay project is to set up a multilingual lexical database for contributions over the web for African languages, notably of Senegal (Wolof, Fula, Bambara, etc.). It must be a basis for the constitution of spell checkers, machine translators, and electronic dictionaries. iBaatukaay seeks to be useful and open to the collaboration of all those who have an interest for the languages concerned and the data generated will be downloadable for free under Creative Commons license.

TËNK

Naka jekk làkki Afrig yi dañu rafle. Li ëpp ci mbéll yi am ak as néew, ci ay këyit lañu leen móol. Jumtukaayu xarala yi am ci làkku Afrig yi lu néew lañu. Saafara yii jafe-jafe moo waral sémbu iBaatukaay. Li yékkati iBaatukaay mooy taxawal ab dàttu baat ñeel i làkk bu ñépp mën a dugal seen loxo ci web ngir làkk Afrig yi, rawatina yoy Senegaal (wolof, pulaar, bàmbara). Warees na cee mën a sukkandiku ngir nas ay jubbantikaayu bind, ay firikaayu làkk ak yeneeni baatukaay. iBaatukaay mên a am njariñ, ku nekk mën cee indi wàllam, rawatina ñi suqali làkk yi soxal ; ñjëriñ li ku nekk mën a cee jot ci mu wut ko jaare ko ci Creative Commons.

MOTS-CLÉS: BASE LEXICALE MULTILINGUE, LANGUES AFRICAINES, SÉNÉGAL, ARCHITECTURE PIVOT, IBAATUKAAY, JIBIKI, XML, WOLOF, PULAAR, BAMBARA, FRANÇAIS.

KEYWORDS: MULTILINGUAL LEXICAL DATABASES, AFRICAIN LANGAGES, SENEGAL, PIVOT ARCHITECTURE, IBAATUKAAY, JIBIKI, XML, WOLOF, FULA, BAMBARA, FRENCH.

BAAT YU CI AM SOLO : DÀTTU BAATIY LÀKK, LAKKI AFRIG, SENEGAAL, DÀTTINU BOOLEY LÀKK, IBAATUKAAY, JIBIKI, XML, WOLOF, PULAAR, BAMBARA, FARAÑSE.

1 Introduction

25 langues endogènes cohabitent au Sénégal avec le français qui est considéré comme la langue officielle du pays. Cependant il convient de faire remarquer que seul 30 % de la population parle le français comparé à certaines langues nationales comme le wolof parlé par 80% de la population.

Malheureusement, les langues nationales du Sénégal comme la plupart des langues africaines n'ont pas bénéficié des avancées du Traitement Automatique du Langage Naturel (TALN) contrairement aux langues européennes. La plupart des ressources qui existent pour ces langues sont en général au format papier. À cela s'ajoute le fait que ces langues sont non enseignées ou trop enseignées.

Le projet iBaatukaay se veut une référence pour les langues africaines notamment sénégalaises. Le but du projet est la conception d'une base lexicale multilingue contributive sur le Web pour les langues africaines notamment sénégalaises de laquelle nous pourrons extraire des dictionnaires destinés à l'enseignement moyen et secondaire mais aussi produire des dictionnaires bilingues (langue locale-langue étrangère et langue locale 1 - langue locale 2). L'aspect collaboratif est important dans la mesure où des contributions sont attendues de toute personne ayant un intérêt pour ces langues à travers le site du projet. Pour un début nous nous focalisons sur les langues sénégalaises suivantes: wolof, pulaar et bambara. Pour mettre en ligne les dictionnaires, nous utiliserons Jibiki (Mangeot et al. 2003), une plate-forme générique en ligne de manipulation de ressources lexicales avec gestion d'utilisateurs et groupes, consultation de ressources hétérogènes et édition générique d'articles de dictionnaires.

Dans la suite de cet article nous aborderons dans un premier temps la problématique du manque de ressources et d'outils TAL pour les langues africaines, dans un deuxième temps nous présenterons le projet, ensuite nous présenterons la méthodologie de transformation des données, enfin nous finirons par une conclusion et donnerons des perspectives.

2 Problématique

2.1 Situation linguistique du Sénégal

Au Sénégal, la reconnaissance des langues nationales est mentionnée dès l'article premier de la constitution du 22 janvier 2001 : «La langue officielle de la République du Sénégal est le Français. Les langues nationales sont le Diola, le Malinké, le Pular, le Sérère, le Soninké, le Wolof et toute autre langue nationale qui sera codifiée».

Il s'avère que le français et le wolof dominent largement dans les transactions langagières. Le Français est parlé par 30% de la population tandis que le wolof est parlé par 80% de la population

(environ 10 millions de locuteurs)1. En plus la langue wolof est une langue véhiculaire au Sénégal et en Mauritanie et parlée en Gambie. La population du Sénégal est à 95% de religion musulmane ce qui fait que certaines langues nationales comme le wolof sont écrites en caractère latin et en Ajami (alphabet arabe complété). Cependant le véritable problème avec les langues africaines en général, en particulier celles parlées au Sénégal c'est que ce sont des langues peu dotées.

2.2 Définition d'une langue peu dotée du point de vue informatique.

C'est un terme utilisé pour désigner le degré d'équipement en outils informatique d'une langue. (Clavier adapté, correcteur orthographique, synthèse de la parole, traducteur automatique, etc.) (Berment, 2004). Ainsi les langues peuvent être classées en 3 groupes: les langues informatiquement peu dotées langues-π (par exemple le wolof, bambara, pulaar, sérère, etc.), les langues moyennement dotées langues-μ (par exemple le portugais, ou le suédois), et langues très bien dotés langues-τ (par exemple, l'anglais, le français).

En effet, en ce qui concerne les langues africaines la plupart des ressources existantes sont en général au format papier. Il existe néanmoins certains travaux concernant les langues africaines notamment sénégalaises. En ce sens nous pouvons citer : le projet de dictionnaire unilingue wolof et bilingue wolof-français (8 167 mots) (Cissé, 2007), le projet DiLAF avec dans ses objectifs un dictionnaire bambara-français (10 800 mots) (Enguehard et al. 2008), ainsi que deux dictionnaires pulaar-français et pulaar-français-anglais du projet ALFFA (African Langages in the field Speech Fundammentals and Automation).

Il existe entre autre pour le wolof un petit corpus sur le Web (60 000 mots), des lexiques du Laboratoire Dynamique du Langage (32 000 mots) ainsi qu'un analyseur morpho-syntaxique (Dione, 2014) mais qui n'a pas encore été testé à grande échelle. Nous reviendrons en détail sur les ressources existantes et leurs caractéristiques dans la partie 3.
Sur le site de Microsoft (http://www.microsoft.com/Language), on y trouve également une banque terminologique Microsoft dans près de 100 langues y compris le wolof. La terminologie est fournie gratuitement sous licence au format .tbx). Dans le projet wikitionary, il existe un dictionnaire multilingue wolof de 2 310 mots qui peut être récupéré. Un dictionnaire bilingue Français-wolof est également disponible sur le site de Glosbe (http://fr.glosbe.com/wo/fr). Un dictionnaire wolof-français est disponible dans le site du projet DilAF (http://pagesperso.lina.univ-nantes.fr/info/perso/permanents/enguehard/DiLAF).

2.3 Motivations

Ce manque de ressources et d'outils de TAL nous motivent à penser qu'une base lexicale multilingue qui servirait de référence pour les langues africaines notamment sénégalaises nous semble très utile d'autant plus qu'elle serait construite de manière contributive ou collaborative sur le Web en utilisant les ressources existantes et les contributions des différents experts de ces langues (lexicologues, lexicographes, linguistes, etc.). Ceci qui nous permettra très rapidement et à moyen terme de pouvoir regrouper tous les mots de chaque langue. En s'appuyant ensuite sur ces bases, des outils tels des analyseurs morphologiques, des correcteurs orthographiques, des corpus, des traducteurs automatiques pourront être développés.

[1]www.francophonie.org/IMG/pdf/repartition_des_francophones_dans_le_monde_en_2014.pdf

3 Présentation du projet iBaatukaay.

Le projet iBaatukaay est un projet dont l'objectif est la conception d'une base lexicale multilingue contributive sur le Web pour les langues africaines notamment sénégalaises. C'est un projet collaboratif. N'importe quel expert du domaine (lexicologues, linguistes, etc.) peut faire des contributions à travers Internet. Les données seront téléchargeable gratuitement depuis la plateforme. Comme cité plus haut, au Sénégal, 25 langues endogènes cohabitent avec le français, l'anglais, l'arabe et les autres langues étrangères. Parmi ces 25 langues nous avons choisi trois langues à savoir le wolof, le pulaar et le bambara dans un premier temps. Le choix n'est pas fortuit. Ce sont des langues largement parlées en Afrique de l'ouest. Le wolof est une langue véhiculaire entre le Sénégal, la Gambie et la Mauritanie. Il est parlé par 10 millions de locuteurs. Le bambara est aussi parlé largement en Afrique de l'ouest par 40 Millions de locuteurs (Gautier & al, 2016). Il est principalement parlé au Mali par 4 millions de locuteurs, au Sénégal, etc. Le pulaar, ou peul ou peulh ou fulfulde, est parlé au Sénégal par 3,5 millions de locuteurs. C'est un dialecte du fula largement parlé en Afrique de l'ouest par 70 millions de locuteurs. Des ressources (dictionnaires au format XML) ont pu être récupérées à travers le projet ALFFA, le projet DiLAF et le projet de dictionnaire de Cissé & al, 2007. Il faut rappeler que toutes ces langues présentent des enjeux pour les multinationales telles que Google et Microsoft. L'interface du moteur de recherche de Google est d'ailleurs traduite en wolof. Le système d'exploitation Windows 8 et les outils de Microsoft (Bing, Outlook, etc.) ainsi leur charte de confidentialité sont disponibles également en wolof, etc.

3.1 Macrostructure de la base lexicale

Pour rappel un dictionnaire 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. La macrostructure d'un dictionnaire représente l'organisation des volumes de ce dictionnaire.

Pour le projet iBaatukaay, nous avons choisi une architecture pivot basée sur la thèse de Gilles Sérasset (Sérasset, 1994), expérimentée à petite échelle dans le projet Papillon (Mangeot, 2001).
Chaque langue du projet sera décrite dans un volume monolingue. Ensuite ces volumes seront reliés entre eux par un volume pivot de liens interlingues appelés acceptions interlingues (axies).
L'architecture pivot est novatrice mais il convient de faire remarquer que scientifiquement elle n'a jamais été testée à grande échelle. Cette hypothèse reste à vérifier et le projet iBaatukaay nous en donne l'occasion. La Figure 1 donne une vue de la macrostructure générale des volumes dans le projet iBaatukaay et la Figure 2 donne une vue détaillée de la macrostructure.

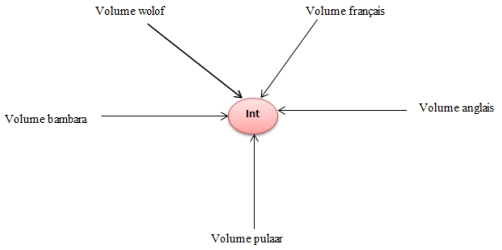


Figure 1: Macrostructure des volumes dans iBaatukaay

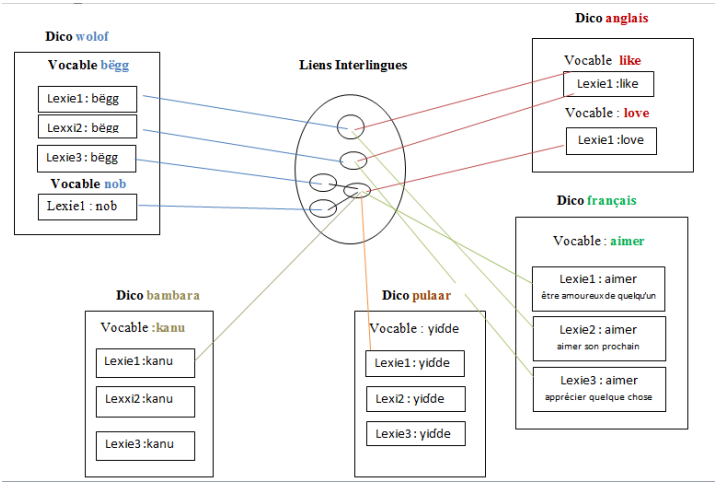


Figure 2: Macrostructure détaillé dans iBaatukaay

3.2 Nomenclature des volumes

Chaque article décrit un mot-forme associé à une catégorie grammaticale. Nous avons décidé de fusionner les vocables homographes de même catégorie grammaticale car les critères pour décider si un mot correspond à un ou plusieurs vocables sont sujets à interprétation.

Par exemple, nous ne distinguerons pas de vocables homographes pour le verbe français « voler ». Il sera l'objet d'un seul article.

Pour le choix des mots qui seront dans le dictionnaire, les critères habituels (existence dans un autre dictionnaire ou dans un corpus) ne peuvent pas être utilisés pour toutes les langues en présence à cause du manque de ressources. Nous nous adapterons donc au cas par cas.

3.3 Microstructure des articles

La structure d'un article constitue la microstructure du dictionnaire. Nous pouvons la considérer comme une structure composée d'objets linguistiques. Dans la microstructure du projet iBaatukaay, chaque article comprend un bloc forme suivi de la catégorie grammaticale du mot-vedette suivi des différents sens du mot-vedette. Dans le bloc forme on trouve le mot-vedette, sa prononciation, ses variantes, la source du mot-vedette et les lexèmes dérivés. Dans chaque bloc sens on a la définition du mot-vedette, la source de la définition, un lien vers l'axie (qui sera reliée aux traductions du mot-vedette dans chaque langue de la base) des liens vers les synonymes, une note d'usage, des exemples en langue locale et la traduction de l'exemple en français. La microstructure est évolutive car nous comptons ajouter la prononciation en utilisant le phonétiseur du projet ALFFA.

3.4 Fonctionnement du projet

Au début, nous allons procéder à la récupération automatique de ressources existantes au format XML.

- Dans le cas où on trouve des fichiers Word, nous adopterons la méthodologie DiLAF (Enguehard et al. 2011).

- Si nous trouvons des dictionnaires imprimés nous adopterons la méthodologie jibiki-Cesselin (Mangeot, 2016).

- Si nous ne trouvons pas de ressources pour une langue donnée, un travail de terrain sera envisagé.

Ensuite, nous nous appuierons sur le Centre Linguistique Appliqué de Dakar (CLAD) à travers ses étudiants pour des contributions en ligne. Pour chaque langue, il faudra nommer un lexicographe en chef, responsable de la validation des articles.

Il convient de faire remarquer que les données produites seront publiquement téléchargeables sous licence de domaine public Creative Commons. Un partenariat est envisagé avec le Ministère de l'éducation nationale et le Ministère de l'enseignement supérieur et de la recherche pour le financement et l'appropriation du projet.

4 Méthodologie de manipulation et de transformation des données

Dans cette partie nous allons parler d'abord des ressources existantes, puis présenter notre méthodologie de transformations des données et enfin présenter les résultats préliminaires obtenus avec le wolof.

4.1 Liste des ressources existantes

4.1.1 Le dictionnaire wolof-français du projet de dictionnaire unilingue wolof et bilingue wolof-français de Cissé & al, 2007.

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 donne comme objectifs principaux :

- 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 Toolbox2 (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. Cela veut dire qu'un vocable polysémique fera l'objet de plusieurs entrées. La Figure 3 présente une illustration d'une entrée ainsi que les champs qui lui sont associés. L'image est obtenue à partir de 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).

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

\lex Lexème wolof	askan
\uttW Transcription phonétique	ɛskɛn
\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 s

Figure 3: Exemple de fiche lexicale obtenu avec l'outil Toolbox

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

4.1.2 Le dictionnaire bambara-français du projet DiLAF

Le projet DiLAF (Dictionnaires Langues Africaines - Français) (Enguehard et al., 2011) vise à convertir des dictionnaires éditoriaux bilingues (bambara, haoussa, kanouri, tamajaq, songhai-zarma) - français en un format XML permettant leur pérennisation et leur partage.

Le dictionnaire éditorial utilisé dans ce projet pour le bambara est le dictionnaire bambara-français du Père Charles Bailleul (édition 1996) comportant 10 000 entrées. Ce dictionnaire est d'abord destiné aux locuteurs francophones désireux de se perfectionner en bambara mais il constitue également une ressource pour les bambaraphones.

Dans ce projet, les vocables homographes de même catégorie grammaticale font l'objet d'une seule entrée dans le dictionnaire. La Figure 4 est un exemple d'entrée de ce dictionnaire.

```
<item id="kanu2">
    <forme>kanu</forme>
    <forme_tons>kànu</forme_tons>
    <cat>v.t.</cat>
    <bloc>
      <sens id="kanu2_1">
        <francais>aimer (parents, amis, amants...)</francais>
      </sens>
      <sens id="kanu2_2">
        <francais>désirer, plaire à... (action)</francais>
        <exemple>
          <ba>u y'a kanu ...</ba>
          <ba_tons>ù y'a kànu ...</ba_tons>
          <fr>il leur a plu de ...</fr>
        </exemple>
      </sens>
    </bloc>
</item>
```

Figure 4 : Exemple d'entrée du dictionnaire : Article Kanu

Il est à noter que le diola, langue reconnue par la constitution du Sénégal, est très proche du bambara. Si nous ne trouvons pas de ressource disponible pour le diola, une solution envisageable serait d'utiliser le dictionnaire bambara-français comme point de départ.

4.1.3 Les dictionnaires fulfulde-français, fulfulde-anglais et fulfulde-français-anglais

Plusieurs dictionnaires existent et ont été convertis dans le cadre des projets DiLAF et ALFFA. Le Tableau 1 suivant donne les caractéristiques de chaque dictionnaire.

Nom du volume	Source	Cibles	Nombre d'entrées
DictionnaireFulNiger_ful_fra	Ful	Fra	4 526
DictionnaireFulUS_eng_ful	Eng	ful	9 997
DictionnaireFulUS_fra_ful	Fra	ful	10 293
DictionnaireFulUS_ful_fra-eng	Ful	Fra eng	10 241

Tableau 1:Caractéristiques des dictionnaires fulfulde récupérés

Pour le volume DictionnaireFulNiger_ful_fra_eng_ful, un article est associé à un mot-vedette en pulaar, suivi de sa catégorie grammaticale, de sa définition et d'un exemple en pulaar, puis d'une traduction en français.

Pour le volume DictionnaireFulUS_eng_ful, un article est associé à un mot-vedette en anglais, suivi de sa prononciation, suivi de sa catégorie grammaticale, et de sa définition en pulaar.

Pour le volume DictionnaireFulUS_fra_ful, un article est associé à un mot-vedette en français, suivi de sa traduction en pulaar.

Pour le volume DictionnaireFulUS_ful_fra-eng, un article est associé à un mot_vedette en pulaar, suivi de sa catégorie grammaticale, de sa définition en pulaar, et de ses traductions en anglais et en français.

4.2 Méthodologie de manipulation et de transformation des données.

L'ensemble des dictionnaires récupérés sont au format XML. Ils sont constitués d'un seul volume bilingue mono-directionnel où on retrouve le **mot-vedette** et sa **traduction** en français.
Pour chaque dictionnaire, il faut passer à une étape de préparation, de tri et de transformation de la microstructure et de la macrostructure pour les convertir vers le format iBaatukaay. Ceci peut être effectué par des scripts PERL.
Cela peut s'avérer lourd si pour chaque dictionnaire on doit écrire des scripts PERL spécifiques pour sa propre transformation.
Ainsi un outil générique de manipulation de dictionnaire XML a été développé dans le cadre du projet. Cet outil nous permet d'effectuer des opérations (Transformation, fusion-interne et réification) sur un ou des dictionnaires au format XML en utilisant les pointeurs CDM obtenus avec iPoLex, un entrepôt de bases lexicales disponible avec la plateforme Jibiki (Zhang et al. 2014).

4.2.1. Présentation d'iPoLex

iPolex est un entrepôt de bases lexicales, accessible par le web. Les interfaces sont programmées en PHP sans connexion à une base de données.
L'ajout d'une nouvelle ressource se fait en trois étapes : d'abord la description de la ressource avec ses métadonnées (langues source et cibles, domaine, auteur, date de création, format, taille des

fichiers XML, etc.) puis, pour chaque volume de la ressource, la description de ses métadonnées (langue source, nombre de mots-vedettes, version, etc.) ainsi que celle des pointeurs CDM (Common Dictionary Markup) (Mangeot, 2002). Ces derniers nous permettent de gérer n'importe quel type de microstructure sans la modifier.

Les pointeurs CDM sont utilisés également pour indexer des parties d'information spécifiques et permettre ensuite une recherche multi-critères. Cette structure est stockée dans un fichier de métadonnées sous forme XML. Les fichiers de métadonnées ont pour but de faciliter l'import de ressources lexicales. Il y a un fichier de métadonnées pour chaque ressource. Dans ce fichier, sont décrites les informations sur cette ressource : les langues source et cibles, l'auteur, les noms et fichiers de volumes, etc. Pour chaque volume, il existe un fichier de métadonnées. Dans ces fichiers, sont décrites toutes les informations des volumes, y compris les pointeurs CDM. Pour chaque pointeur CDM, on indique le chemin XPath vers l'élément correspondant dans la microstructure XML. Voir la figure 5 - « Exemple de CDM ».

La description des ressources sur iPolex se termine enfin par la création d'un répertoire sur le serveur et la génération de fichiers de métadonnées.

La dernière étape est celle du téléversement des fichiers de données sur le serveur. Pour cela, l'entrepôt peut se monter comme répertoires distant grâce au protocole WebDAV.

```
<cdm-entry xpath="/database/lexGroup"/>
<cdm-entry-id xpath="/database/lexGroup/@id"/>
<cdm-headword xpath="/database/lexGroup/lex/text()"/>
<cdm-headword-variant xpath="/database/lexGroup/varW/text()"/>
<cdm-pronunciation xpath="/database/lexGroup/uttW/text()"/>
<cdm-pos xpath="/database/lexGroup/catWGroup/catW/text()"/>
<cdm-definition xpath="/database/lexGroup/defWGroup/defW/text()"/>
<cdm-translation xpath="/database/lexGroup/tradFlexGroup/tradFlex/text()" d:lang="fra" />
<cdm-example-block xpath="/database/lexGroup/phrWGroup"/>
<cdm-example xpath="/database/lexGroup/phrWGroup/tradPhrW/text()" d:lang="fra" />
```

Figure 5: Extrait Pointeurs CDM à partir du fichier de métadonnées du dico Cissé & al (Voir Figure 3) après ajout sur iPolex

4.2.2. Présentation de l'outil de manipulation et de transformation des données

L'outil permet de faire de faire des opérations sur des dictionnaires au format XML telles que la transformation d'un dictionnaire d'une microstructure source vers une microstructure cible, la fusion interne d'un dictionnaire mais aussi la réification d'un dictionnaire. La fusion interne est la fusion des vocables homographes de même catégorie grammaticale. La réification permet de générer un dictionnaire source, un dictionnaire cible et un dictionnaire pivot. Il convient de rappeler dans pour le projet iBaatukaay, nous avons opté pour une macrostructure pivot. La figure 6 représente la microstructure cible de toutes les ressources récoltées dans le cadre du projet.

Chaque dictionnaire récolté dans le cadre du projet est ajouté sur iPolex afin d'avoir une vue CDM du dictionnaire (fichier de métadonnées du dictionnaire contenant les pointeurs CDM, dictionnaire au format brut,..). Nous ajouterons aussi sur iPolex un dictionnaire appelé DicoArrivee avec comme unique entrée notre microstructure cible ou notre template (voir Figure 6 et 7).

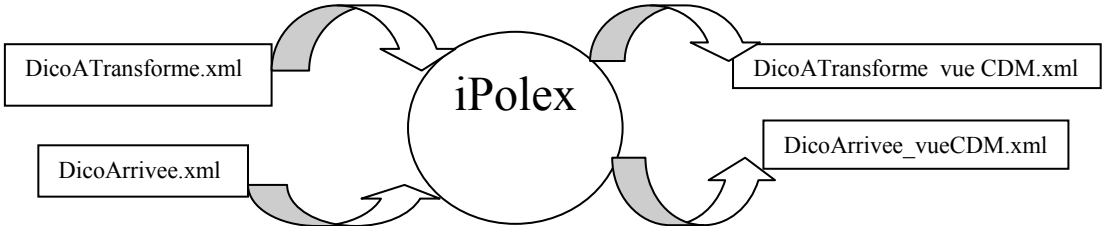


Figure 7:Ajout Dictionnaire sur iPolex

```
<article id="">
 <bloc_forme>
  <mot_vedette/>
  <source_mot_vedette/>
  <lexème_source/>
  <variante/>
 <prononciation/>
 </bloc_forme>
 <catégorie_grammaticale/>
 <classe_nominale/>
 <bloc_sens>
  <sens id="">
  <définition/>
  <source_définition/>
  <bloc_traduction>
   <traduction_française/>
   <catégorie_grammaticale_traduction_française_mot_vedette/>
   <lien_traduction d:lang="" id="" type="" volume=""/>
  </bloc_traduction>
  <exemples>
  <exemple>
   <exemple-wol/>
   <exemple-fra/>
  </exemple>
  </exemples>
  <synonyme/>
  <homonyme/>
  <note_usage/>
 <bloc_dérivés>
  <expression_dérivée/>
 </bloc_dérivés>
</sens>
 </bloc_sens>

 <sources>
 <entree-source provenance="" />
 </sources>
```

Figure 8:microstruture cible de nos dictionnaires cible ou template

4.2.2. Présentation de l'algorithme de transformation.

La transformation nécessite la préparation d'abord des données. La préparation consiste à couper l'entête et le pied de page du volume XML. L'algorithme de transformation prend en entrées un dictionnaire source, le fichier de métadonnées du dictionnaire, le fichier de métadonnées de notre dictionnaire d'arrivée et le fichier de template et produit en sortie un dictionnaire cible respectant la microstructure iBaatukaay (Voir Figure 8). Les figures 9 et 10 représentent l'article « **aada** » avant et après transformation pour le dictionnaire (Cissé & al, 2007). Pour éviter des pertes des données des dictionnaires sources, nous avons créé dans les dictionnaires cibles une balise **entree-source** qui permet de garder l'article au format d'origine ainsi que sa provenance.

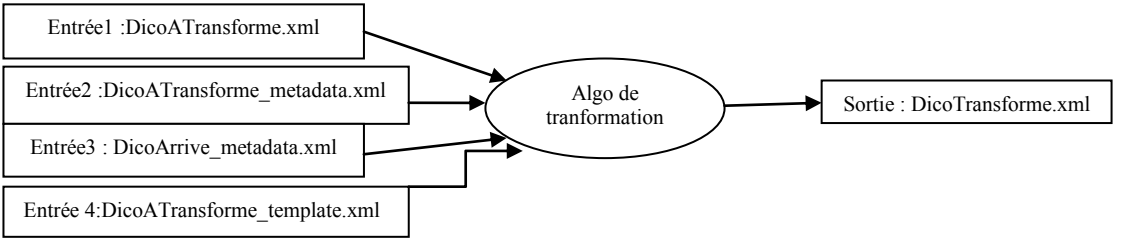


Figure 9: Arguments d'entrées et de sortie de l'algorithme de transformation

```
<article id="aada1" level="">
  <bloc_forme>
    <mot_vedette>aada</mot_vedette>
    <prononciation>a:de</prononciation>
  </bloc_forme>
  <catégorie_grammaticale>turu bokkaale</catégorie_grammaticale>
  <classe_nominale>j-</classe_nominale>
  <sens id="aada1_1">
    <définition>li aw xeet cosaanoo di ko def</définition>
    <bloc_traduction>
      <traduction_française>coutumes</traduction_française>
      <catégorie_grammaticale_traduction_française_mot_vedette>nom
      </catégorie_grammaticale_traduction_française_mot_vedette>
    </bloc_traduction>
    <exemple>
      <phrase_illustration>Sunu aada day bañ foot àllarba</phrase_illustration>
      <traduction_française_phrase_illustration>Nos coutumes nous interdisent
      de faire le linge le mercredi
      </traduction_française_phrase_illustration>
    </exemple>
  </sens>
  <bloc_métainformation>
    <auteur>MTC</auteur>
    <date_dernière_modification>02/Sep/2007</date_dernière_modification>
  </bloc_métainformation>
</article>
```

Figure 10:Article aada avant transformation (Dictionnaire Source Cissé & al, 2007)

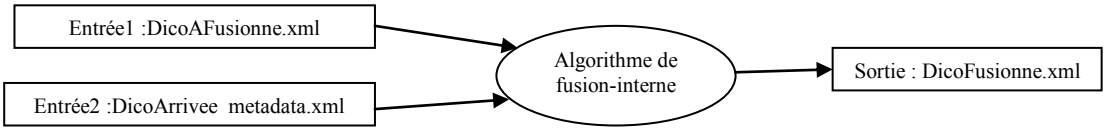
```
<article id="">
   <bloc_forme>
              <mot_vedette>aada</mot_vedette>
             <source_mot_vedette/>
             <lexème_source/>
            <variante/>
           <prononciation>a:dɐ</prononciation>
 </bloc_forme>
 <catégorie_grammaticale>nom</catégorie_grammaticale>
 <classe_nominale>j-</classe_nominale>
  <bloc_sens>
    <sensid="">
      <définition>li aw xeet cosaanoo di ko def</définition>
       <source_définition/>
      <bloc_traduction>
        <traduction_française>coutumes</traduction_française>
<catégorie_grammaticale_traduction_française_mot_vedette>nom</catégorie_grammaticale_traduction_franç
aise_mot_vedette>
       <lien_traduction d:lang="" id="" type="" volume=""/>
    </bloc_traduction>
     <exemples>
        <exemple>
          <exemple-wol>Sunu aada day bañ foot àllarba</exemple-wol>
          <exemple-fra>Nos coutumes nous interdisent de faire le linge le mercredi</exemple-fra>
        </exemple>
     </exemples>
     <synonyme/>
     <homonyme/>
     <note_usage/>
     <bloc_dérivés>
      <expression_dérivée/>
      </bloc_dérivés>
      </sens>
   </bloc_sens>
  <sources>
<entree-source provenance="Thierno"><lexGroup>  <lex>aada</lex>  <uttW>a:dɐ</uttW>  <catWGroup>
<catW>turu bokkaale</catW>  <clasW>j-</clasW>  </catWGroup>  <defWGroup>  <defW>li aw xeet
cosaanoo di ko def</defW>  </defWGroup>  <tradFlexGroup>  <tradFlex>coutumes</tradFlex>
<catF>nom</catF>  </tradFlexGroup>  <phrWGroup>  <phrW>Sunu aada day bañ foot àllarba</phrW>
<tradPhrW>Nos coutumes nous interdisent de faire le linge le mercredi</tradPhrW>  </phrWGroup>
<aut>MTC</aut>  <dat>02/Sep/2007</dat>  </lexGroup></entree-source>
</sources>
</article>
```

Figure 11: Article **aada** après transformation

4.2.3. Présentation de l'algorithme de fusion-interne

La fusion-interne se déroule en 3 étapes:

1. La préparation : Elle consiste à couper l'entête et le pied de page du volume XML.
2. Le tri des articles du dictionnaire XML selon l'ordre alphabétique.
3. La fusion des vocables homographes de même catégorie grammaticale et la création des sens de mot équivalents.



La Figure 11 et 12 représente deux entrées homographes de même catégorie grammaticale dans le Dictionnaire de Cissé & al. La Figure 13 donne le résultat des deux entrées fusionnées en appliquant l'algorithme de fusion.

```
<article id="">
    <bloc_forme>
      <mot_vedette>aloom</mot_vedette>
      <source_mot_vedette/>
      <lexème_source/>
      <variante/>
      <prononciation>elɔ:mˋ</prononciation>
    </bloc_forme>
    <catégorie_grammaticale>nom</catégorie_grammaticale>
    <classe_nominale>g-</classe_nominale>
   <bloc_sens>
    <sens id="">
     <définition>garab giy meññ aloom</définition>
     <source_définition/>
     <bloc_traduction>
       <traduction_française>diospyros</traduction_française>
       <catégorie_grammaticale_traduction_française_mot_vedette>nom</catégorie_grammaticale_
       <lien_traduction d:lang="" id="" type="" volume=""/>
     </bloc_traduction>
     <exemples>
     <exemple>
       <exemple-wol>Gor nañu aloom gi nekkoon ci peñc mi.</exemple-wol>
       <exemple-fra>On a coupe le diospyros qui était sur la place.</exemple-fra>
     </exemple>
     </exemples>
     <synonyme/>
     <homonyme>aloom</homonyme>
     <note_usage/>
    <bloc_dérivés>
     <expression_dérivée/>
    </bloc_dérivés>
   </sens>
    </bloc_sens>
    <sources>
     <entree-source provenance="Thierno"><lexGroup>   <lex>aloom</lex>   <uttW>elɔ:mˋ</uttW>
    </sources>
  </article>
```

Figure 12:Entrée1: aloom (Dico Cissé & al)

```
<article id="">
    <bloc_forme>
      <mot_vedette>aloom</mot_vedette>
      <source_mot_vedette/>
      <lexème_source/>
      <variante/>
      <prononciation>ɛlɔ:mˋ</prononciation>
    </bloc_forme>
    <catégorie_grammaticale>nom</catégorie_grammaticale>
    <classe_nominale>b-</classe_nominale>
   <bloc_sens>
    <sens id="">
     <définition>xeetu meñent mu ñu mëna lekk</définition>
     <source_définition/>
     <bloc_traduction>
       <traduction_française>fruit du Diospyros</traduction_française>
       <catégorie_grammaticale_traduction_française_mot_vedette>nom</catégorie_grammaticale_tra
       <lien_traduction d:lang="" id="" type="" volume=""/>
     </bloc_traduction>
     <exemples>
     <exemple>
       <exemple-wol>Xale bi dafa fori aloom.</exemple-wol>
       <exemple-fra>L'enfant est parti ramasser des fruits de diospyros.</exemple-fra>
     </exemple>
     </exemples>
     <synonyme/>
     <homonyme>aloom</homonyme>
     <note_usage/>
    <bloc_dérivés>
     <expression_dérivée/>
    </bloc_dérivés>
   </sens>
    </bloc_sens>
    <sources>
     <entree-source provenance="Thierno"><lexGroup>   <lex>aloom</lex>   <uttW>ɛlɔ:mˋ</uttW>   <
    </sources>
  </article>
```

Figure 13:Entrée 2:aloom (Dico Cissé & al)

```
<article id="">  <bloc_forme>  <mot_vedette>aloom</mot_vedette>  <source_mot_vedette/>  <lexème_source/>  <variante
/>  <prononciation>ɛlɔ:mˋ</prononciation>  </bloc_forme>  <catégorie_grammaticale>nom</catégorie_grammaticale>  <
lasse_nominale>b-</classe_nominale>  <bloc_sens>  <sens id="s1">  <définition>xeetu meñent mu ñu mëna lekk</définition>
<source_définition/>  <bloc_traduction>  <traduction_française>fruit du Diospyros</traduction_française>  <
atégorie_grammaticale_traduction_française_mot_vedette>nom</catégorie_grammaticale_traduction_française_mot_vedette>  <
lien_traduction d:lang="" id="" type="" volume=""/>  </bloc_traduction>  <exemples>  <exemple>  <exemple-wol>Xale
i dafa fori aloom.</exemple-wol>  <exemple-fra>L'enfant est parti ramasser des fruits de diospyros.</exemple-fra>  </exemple>
</exemples>  <synonyme/>  <homonyme>aloom</homonyme>  <note_usage/>  <bloc_dérivés>  <
xpression_dérivée/>  </bloc_dérivés>  </sens><sens id="s2">  <définition>garab giy meññ aloom</définition>  <
ource_définition/>  <bloc_traduction>  <traduction_française>diospyros</traduction_française>  <
atégorie_grammaticale_traduction_française_mot_vedette>nom</catégorie_grammaticale_traduction_française_mot_vedette>  <
lien_traduction d:lang="" id="" type="" volume=""/>  </bloc_traduction>  <exemples>  <exemple>  <exemple-wol>Gor
ñu aloom gi nekkoon ci peñc mi.</exemple-wol>  <exemple-fra>On a coupe le diospyros qui était sur la place.</exemple-fra>
</exemple>  </exemples>  <synonyme/>  <homonyme>aloom</homonyme>  <note_usage/>  <bloc_dérivés>  <
xpression_dérivée/>  </bloc_dérivés>  </sens>  </bloc_sens>  <sources>  <entree-source provenance="Thierno"><lexGroup>
<lex>aloom</lex>  <uttW>ɛlɔ:mˋ</uttW>  <catWGroup>  <catW>turu bokkaale</catW>  <clasW>b-</clasW>  </catWGroup>  <srclW>
ictionnaire arame</srclW>  <defWGroup>  <defW>xeetu meñent mu ñu mëna lekk</defW>  </defWGroup>  <homW>aloom</homW>  <tradFlexGroup>
<tradFlex>fruit du Diospyros</tradFlex>  <catF>nom</catF>  </tradFlexGroup>  <phrWGroup>  <phrW>Xale bi dafa fori aloom.</phrW>  <
radPhrW>L'enfant est parti ramasser des fruits de diospyros.</tradPhrW>  </phrWGroup>  <aut>MTC</aut>  <cmt>décrire le fruit dans la
ef wol; meñnent?</cmt>  <autStat>AMD</autStat>  <dat>31/Aug/2007</dat>  </lexGroup></entree-source>  <entree-source
rovenance="Thierno"><lexGroup>  <lex>aloom</lex>  <uttW>ɛlɔ:mˋ</uttW>  <catWGroup>  <catW>turu bokkaale</catW>  <clasW>g-</clasW>  </
atWGroup>  <srclW>dictionnaire arame</srclW>  <defWGroup>  <defW>garab giy meññ aloom</defW>  </defWGroup>  <homW>aloom</homW>  <
radFlexGroup>  <tradFlex>diospyros</tradFlex>  <catF>nom</catF>  </tradFlexGroup>  <phrWGroup>  <phrW>Gor nañu aloom gi nekkoon ci
peñc mi.</phrW>  <tradPhrW>On a coupe le diospyros qui était sur la place.</tradPhrW>  </phrWGroup>  <aut>MTC</aut>  <cmt>meññal?
meññil?</cmt>  <autStat>AMD</autStat>  <dat>28/Aug/2007</dat>  </lexGroup></entree-source>  </sources>  </article>
```

Figure 14:Résultat de la fusion des deux entrées en utilisant l'algorithme de fusion

4.2.3. Présentation de l'algorithme de réification

La réification consiste à partir d'un volume bilingue monodirectionnel, de générer trois volumes (un volume source, un volume cible et un volume pivot qui sert de liens interlingues entre le volume source et le volume cible).

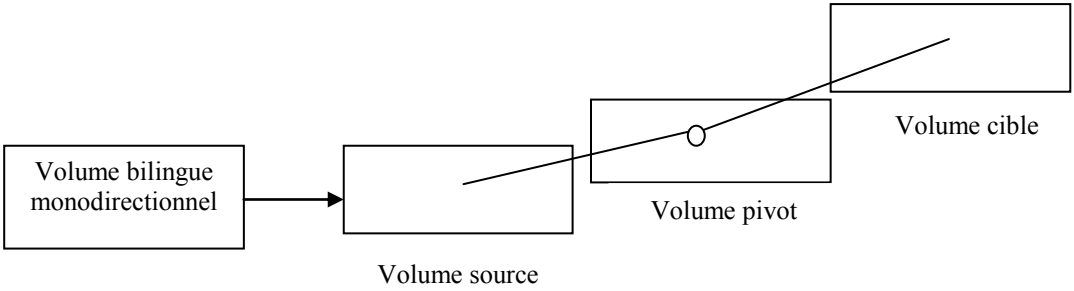


Figure 15: Architecture pivot

L'étape de réification se déroule en trois étapes :

1. L'identification des articles.
2. La création du volume source et volume cible avec les liens de traduction entre des ces deux volumes.
3. La réification proprement dite qui consiste à créer le volume pivot qui sert de liens interlingues entre les deux volumes.

L'article **aada** de la figure 10 réifié avec notre algorithme donne dans le volume cible la figure 16 Le lien de traduction nous permettra de trouver sa traduction dans le volume axi (Figure 17). La figure 18 est la traduction du mot en **aada** en français qui signifie **coutume**.

```
<article id="1">
    <bloc_forme>
      <mot_vedette>aada</mot_vedette>
      <source_mot_vedette/>
      <lexème_source/>
      <variante/>
      <prononciation>a:dɒ</prononciation>
    </bloc_forme>
    <catégorie_grammaticale>nom</catégorie_grammaticale>
    <classe_nominale>j-</classe_nominale>
   <bloc_sens>
     <sens id="">
      <définition>li aw xeet cosaanoo di ko def</définition>
      <source_définition/>
      <bloc_traduction>
        <lien_traduction d:lang="axi" id="axi.[wol:1,fra:coutumes].1.e" type="pivot" volume="DicoArrivee_axi"/>

        <lien_traduction d:lang="" id="" type="" volume=""/>
      </bloc_traduction>
      <exemples>
      <exemple>
        <exemple-wol>Sunu aada day bañ foot àllarba</exemple-wol>
        <exemple-fra>Nos coutumes nous interdisent de faire le linge le mercredi</exemple-fra>
      </exemple>
      </exemples>
      <synonyme/>
      <homonyme/>
      <note_usage/>
     <bloc_dérivés>
      <expression_dérivée/>
    </bloc_dérivés>
   </sens>
    </bloc_sens>
  </article>
```

Figure 16:Article **aada** avec son lien de traduction dans le volume wolof

```
<axie id="axi.[wol:1,fra:coutumes].1.e">
  <mot_vedette>[wol:1,fra:coutumes]</mot_vedette>
  <reflexies>
    <reflexie idarticle="" idsens="" d:lang="" volume="" type=""/>
  <reflexie idarticle="1" idsens="" d:lang="wol"
  volume="DicoArrivee_wol_fra" type="final"/>
  <reflexie idarticle="fra.coutumes.1.e" idsens="" d:lang="fra" volume="DicoArrivee_fra" type="final"/>
  </reflexies>
</axie>
```

Figure 17:Article dans le volume qui permet de trouver la traduction du mot **aada**

```
<article id="fra.coutumes.1.e">
    <bloc_forme>
      <mot_vedette>coutumes</mot_vedette>
      <source_mot_vedette/>
      <lexème_source/>
      <variante/>
      <prononciation/>
    </bloc_forme>
    <catégorie_grammaticale>nom</catégorie_grammaticale>
    <classe_nominale/>
   <bloc_sens>
     <sens id="">
      <définition/>
      <source_définition/>
      <bloc_traduction>
        <traduction_française/>
        <catégorie_grammaticale_traduction_française_mot_vedette/>
        <lien_traduction d:lang="axi" id="axi.[wol:1,fra:coutumes].1.e" type="pivot" volume="DicoArrivee_axi"/>
      </bloc_traduction>
      <exemples>
      <exemple>
        <exemple-wol/>
        <exemple-fra/>
      </exemple>
      </exemples>
      <synonyme/>
      <homonyme/>
      <note_usage/>
     <bloc_dérivés>
      <expression_dérivée/>
    </bloc_dérivés>
   </sens>
    </bloc_sens>
    <sources>
     <entree-source provenance=""/>
    </sources>
  </article>
```

Figure 18: Article **coutume** qui est la traduction du mot **aada** dans volume cible (français)

5 Conclusion et perspectives

Les langues du Sénégal comme la plupart des langues africaines nécessitent d'être outillées pour leur visibilité sur la toile et leur insertion dans le système académique.
D'où le projet iBaatukaay qui est un projet dont la finalité est de mettre en place une base lexicale multilingue à structure pivot contributive sur le Web, qui pourra servira de modèle pour les langues du Sénégal. Nous nous appuierons sur le CLAD à travers ses étudiants pour la contribution en ligne et la vérification des données, Dans nos futurs travaux nous comptons :

- mettre les données sur jibiki en respectant la macrostructure d'iBaatukaay (Architecture pivot);
- ouvrir les contributions en ligne;
- vérifier l'hypothèse si jibiki pourra tenir si l'architecture pivot passe à grande échelle ; dans le cas contraire identifier, les problématiques de recherche que cela va soulever;
- convertir chaque dictionnaire monolingue au format LMF (Lexical Markup Framework) ;
- utiliser l'analyseur morphologique du wolof développé par Cheikh Bamba Dione (2012) comme lemmatiseur pour faire ce qu'on appelle de la lecture active pour le wolof dans le projet iBaatukaay ;

- implémenter des analyseurs morphologiques pour le pulaar et les autres langues ;
- utiliser ces analyseurs pour en faire des correcteurs orthographiques ;
- implémenter des corpus pour chaque langue nationale ;
- programmer des outils de traduction automatique.

Remerciements

Nous remercions le Centre d'Excellence Africain en Mathématiques, Informatique et TIC d'avoir soutenu le projet et d'avoir financé à un des auteurs du projet quatre mois de séjour de recherche au Laboratoire LIG de Grenoble.

Références

Berment V. Méthodes pour informatiser des langues et des groupes de langues ”peu dotées”. Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, Grenoble, France, 2004.Formal and Computational Aspects of Wolof Morphosyntax in Lexical Functional Grammar.

Chalvin A., Mangeot M. Méthodes et outils pour la lexicographie bilingue en ligne : le cas du Grand Dictionnaire Estonien-Français. Actes d'EURALEX 2006, Turin, Italie, 6-9 septembre, 2006.

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.

Dione C.B. (2014). Thèse de Nouveau Doctorat. Université de Bergen. Norvège.

Dione C.B. (2012). A Morphological Analyzer For Wolof Using Finite-State Techniques. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry De clerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.), Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC’12), Istanbul, Turkey: ELRA.

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

Gauthier E., Besacier L., Voisin S., Melese M., Elingui U. P. (2016) Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof. LREC’2016

Mangeot M. Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue. Thèse de nouveau doctorat, spécialité informatique, Université Joseph

MANGEOT M. (2002). An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language, in Proc. of Papillon 2002 Workshop (www.papillon-dictionary.org/static/info_media/1620673.pdf)

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.

Mangeot M. (2016) Collaborative construction of a good quality broad coverage and copyright free Japanese-French dictionary Hosei University International Found Foreign Scholar Fellowship Report Volume XVI 2013-2014, Hosei University, Tokyo, Japan, pp. 175-208

Mangeot M. Projet MotÀmot : élaboration d'un système lexical multilingue par le biais de dictionnaires bilingues. Actes des journées scientifiques LTT 2009, Lisbonne, Portugal, 15-17 octobre, 12 p., 2009.

Sérasset G. (1994) SUBLIM: un Système Universel de Bases Lexicales Multilingues et NADIA: sa spécialisation aux bases lexicales interlingues par acceptions. Thèse de nouveau doctorat, Spécialité Informatique, Université Joseph Fourier Grenoble 1, 194 p.

Zhang Y., Mangeot M., Bellynck V., Boitet Ch. (2014) Jibiki-LINKS: a tool between traditional dictionaries and lexical networks for modelling lexical resources. Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex) 2014 (Eds. Michael Zock, Reinhard Rapp, Chu-Ren Huang), Dublin, Ireland, 23 August 2014, 12 p.