

## **SUJET DE THESE : Mise en place d'un dictionnaire collaboratif basé sur LMF intégré au web sémantique**

**Résumé du projet de thèse :** Dans plusieurs pays francophones d'Afrique de l'Ouest, les populations dans leur majorité ne comprennent pas le français, langue officielle et qui permet d'accéder aux savoirs et aux formations, gage d'un développement socio-économique réel et durable.

C'est le cas au Sénégal où le français, langue officielle du pays, n'est comprise que par 15 à 20% de la population [6][7]. Au même moment, cette même population, à plus de 80%, utilise dans ses activités au quotidien, le wolof, langue endogène et véhiculaire.

Dans la plupart de ces pays, Sénégal inclus, les langues endogènes sont malheureusement peu dotées d'outils modernes comme des dictionnaires normalisés et des correcteurs interactifs, nécessaires à leur introduction efficace dans le système éducatif pour accompagner le français dans la prise en charge des besoins en formation des populations.

C'est dans cette optique que l'équipe TALN du LANI (Laboratoire d'Analyse Numérique et d'Informatique) de l'UGB a mis en place et a pour ambition d'exécuter le plan de recherche suivant :

1/ Mettre en place une base lexicale multi-usages pour le wolof en construction collaborative sur le Web. Il s'agit de partir de la base de données issues du projet (cissé et al) [2] pour mettre en place un dictionnaire collaboratif en ligne répondant à la norme LMF [3] où les acteurs (linguistes, lexicologue, lexicographe, chercheur, etc.) auront la possibilité d'utiliser et de mettre à jour la base de données. Ce qui permet d'avoir les ressources nécessaires pour la suite.

2/ Mettre en place un dictionnaire collaboratif basé sur LMF intégré au web sémantique. Ceci en construisant une dimension sémantique pour la ressource décrite au point 1/, qui permette son exploitation au niveau du web sémantique par le public comme par les services web, que ce soit en l'intégrant avec d'autres ressources, en réutilisant ces données ou en concevant des services innovants à partir de telles données.

3/ Mettre en place un correcteur orthographique interactif pour la langue Wolof en partant des résultats des points 1 et 2/.

4/ Mettre en place un correcteur grammatical interactif pour la langue Wolof

5/ Mettre en place un outil de traduction automatique.

6/ Mettre en place un outil de reconnaissance vocale et un outil de synthèse vocale.

7/ etc.

Le point 1/ du plan de recherche est en cours de finalisation à travers la thèse de Mouhamadou Khoulé dont le sujet est « *Étude et mise en œuvre d'une base lexicale multi-usages pour le wolof en construction collaborative sur le Web.* ». L'équipe TALN-UGB collabore dans cette phase avec d'éminents chercheurs dont Mathieu Mangeot chercheur du TALN spécialiste des dictionnaires électroniques à l'Université de Grenoble et co-encadrant de M. Khoulé, Mame Thierno Cissé Linguistique-Phonétique au département de Linguistique et Sciences du Langage, FLSH, UCAD et Sokhna Bao Diop linguiste à l'UGB.

Notons qu'il est prévu avant la fin de cette thèse, dont la soutenance doit avoir lieu avant décembre 2016, le développement d'un générateur de formes fléchies à partir du dictionnaire en prélude à un

nouveau sujet de thèse dont l'objectif est de mettre en place un correcteur orthographique pour la langue wolof.

L'objectif de ce **présent** projet de thèse est de réaliser le point 2/. Cela consiste à transformer la base lexicale issue du point 1/ en un modèle permettant de partager et de réutiliser les données entre plusieurs applications, entreprises et groupes d'utilisateurs.

Le projet de thèse de Mouhamad Ndiankho Thiam dont le sujet est « **Mise en place d'un dictionnaire collaboratif basé sur LMF intégré au web sémantique** », rentre dans ce cadre. Nous collaborons dans cette phase avec d'éminents chercheurs dont Béatrice Bouchou Markhoff chercheur spécialiste en bases de données et traitement des langues naturelles et co-encadrant de M. Thiam, Denis Maurel informaticien-linguiste à l'Université de François Rabelais de Tours, Mathieu Mangeot chercheur du TALN spécialiste des dictionnaires électroniques à l'Université de Grenoble, Mame Thierno Cissé Linguistique-Phonétique au département de Linguistique et Sciences du Langage, FLSH, UCAD.

### **Contexte et problématique**

Au Sénégal comme dans presque tous les pays africains, la majeure partie de la population ne comprend pas la langue officielle qui est souvent une langue étrangère héritée de la colonisation. Cette situation constitue un véritable frein au développement socio-économique du pays si nous n'utilisons pas, comme alternatives, les langues nationales comme le wolof (compris et parlé par plus de 80% de la population d'après le recensement de 2008).

Comment peut-on développer un pays avec seulement une langue que la majorité de la population ne comprend pas ?

Rappelons qu'une langue est un système de signes linguistiques, vocaux, graphiques ou gestuels, qui permet la communication entre des individus. Il faut savoir qu'aucune langue n'est née écrite. Toute langue parlée peut être conventionnellement écrite et enrichie avec des mots d'autres langues. L'essentiel est que les populations puissent l'utiliser pour bien communiquer et pour améliorer leurs conditions de vie. Tel n'est pas le cas du français au Sénégal utilisé dans l'administration, l'enseignement, les médias et le monde des affaires, les panneaux d'indication, les panneaux publicitaires, etc. Même si cette langue n'est comprise que par 20% à 30% de la population (d'après le dernier rapport de la francophonie).

Utiliser nos langues locales présente plusieurs avantages économiques et socio-culturels. Nombre d'emplois pourraient être créés en traduisant en wolof des centaines ou même des milliers de livres, de bandes dessinées, de logiciels et de jeux écrits dans d'autres langues, voire en interprétant en wolof des films, des documentaires, des dessins animés et des théâtres produits dans d'autres langues à travers le monde. Et cela aiderait une frange importante de la population qui a de plus en plus accès aux outils numériques d'accéder facilement à des contenus adaptés à leurs besoins. Ne pas utiliser nos langues nous prive tout simplement de ces possibles créations d'emplois que nos jeunes tentent d'aller chercher en Occident.

Ce projet va contribuer à lever quelques contraintes sur le wolof et par extension sur nos différentes langues pour leur permettre d'influer positivement sur l'amélioration des conditions de vie de nos populations.

## **Objectifs du projet de thèse**

### ***Objectif général :***

Avoir un dictionnaire wolof collaboratif interopérable.

### ***Objectifs spécifiques :***

- Définition d'un modèle de dictionnaire adapté au Wolof et basé sur LMF
- Construction d'un dictionnaire collaboratif basé sur ce modèle
- Intégration de la ressource au Web Sémantique
- Enrichissement automatiquement par d'autres ressources existantes
- Expérimentation de différents cas d'utilisation

## **Encadrement :**

El Hadji Mamadou NGUER, Ousmane THIARE; Laboratoire d'Analyse Numérique et d'Informatique (LANI)

El hadji Mamadou Nguer qui est le coordonnateur de ce projet, y contribuera principalement :

- sur les modèles de données du web, que ce soit XML ou les standards du web sémantique (RDF, RDFS et OWL)
- par son expertise sur la langue wolof et son expérience sur le dictionnaire sur le quel ce travail va s'appuyer. Il pourrait par la suite, dans le cadre du projet global, participer à l'implantation du wolof dans le logiciel libre Unitex et apporter son expertise pour la validation des formes fléchies générées dans le cadre de ce logiciel. Il pourrait aussi contribuer à une version wolof de Prolex et de CasEN.

## **Collaborateurs :**

Mame Thierno CISSE est spécialiste de Phonétique et ces principaux travaux dans le domaine portent sur le wolof. Par ailleurs depuis 2004 il s'est intéressé aux questions de lexiques et de traduction ce qui l'a amené à initier et à coordonner le projet de « conception d'un dictionnaire électronique unilingue wolof et bilingue wolof-français » en partenariat avec l'Institut für Linguistik-Phonetik de l'Université de Cologne et l'Institut supérieur de traducteurs et interprètes de Bruxelles (Centre de recherche Termisti).

Béatrice Bouchou Markhoff travaille dans l'équipe BdTIn ("Bases de données et Traitement automatique des langues naturelles") du LI (Laboratoire d'Informatique - EA 6300) de l'Université François Rabelais de Tours, sur les modèles de données du web : XML (schémas et contraintes d'intégrité, évolution, à base d'automates d'arbres et grammaires attribuées), et web sémantique (ontologies, médiation sémantique, logiques de description). Ses champs d'application sont le TAL (représentation de ressources lexicales comme les noms propres, en LMF) et le patrimoine culturel (projets liés à la Renaissance, l'archéologie, les manuscrits de l'Ouest Saharien).

Denis Maurel (**université François-Rabelais de Tours**) est informaticien, spécialiste du traitement automatique des langues, avec option linguistique.

## Références

1. 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.
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. Dictionnaires wolof en ligne : état de l'art et perspectives. E. M. Nguer, M. Khoulé, M. D. Thiam. M. B. THIAM, O. THIARE, M. T. CISSE, M. MANGEOT. CNRIA 2015.
5. Vers la mise en place d'un lexique basé sur LMF pour la langue wolof. E. M. Nguer, M. Khoulé, M. D. Thiam. TALN-RECITAL 2014 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.
6. Rapport 2014 de la répartition des francophones dans le monde [http://www.francophonie.org/IMG/pdf/repartition\\_des\\_francophones\\_dans\\_le\\_monde\\_en\\_2014.pdf](http://www.francophonie.org/IMG/pdf/repartition_des_francophones_dans_le_monde_en_2014.pdf)
7. Rapport 2013 de la banque mondiale sur la population mondiale <http://data.worldbank.org/indicator/SP.POP.TOTL>

**Profil recherché :** Titulaire d'un Master en Informatique