MISE AU POINT D'UNE BASE DE DONNÉES LEXICALE MULTIFONCTIONNELLE : LE DICTIONNAIRE UNILINGUE WOLOF ET BILINGUE WOLOF-FRANÇAIS

Mame Thierno Cissé, Anna Marie Diagne
Université Cheikh Anta Diop (Dakar, Sénégal)
Marc van Campenhoudt, Paul Muraille
Centre de Recherche Termisti (ISTI, Bruxelles, Belgique)

à paraître dans les actes des
5es Journées de linguistique de corpus
(Lorient, 13 - 15 septembre 2007)

1. INTRODUCTION

Étalé sur une période de deux ans (2007-2009), le projet de mise au point d'une base de données lexicale multifonctionnelle est un projet mené par l'Université Cheikh Anta Diop de Dakar au Sénégal en collaboration avec le centre de recherche en linguistique appliquée Termisti de la Haute École de Bruxelles (Belgique) et avec le soutien de l'Agence universitaire de la francophonie (AUF). Ce projet, aujourd'hui à mi-parcours, a pour objectif principal de constituer une base de données lexicale multifonctionnelle pour la langue wolof, en d'autres termes, de collecter, numériser et standardiser un matériel lexical qui pourra être exploité à une double fin : d'une part, fournir aux chercheurs et aux spécialistes du wolof un corpus lexical réexploitable à des fins d'enrichissement du matériel constitué dans le cadre du projet ou réutilisable pour des applications de l'ingénierie linguistique ; d'autre part, fournir le support d'un dictionnaire à destination des populations wolophones monolingues ou bilingues.

Structuration et format des données ont été pensés en premier lieu pour que le produit final puisse servir au double usage de corpus lexical et de dictionnaire (tantôt à l'usage exclusif de l'une ou l'autre de ces fonctions tantôt à l'usage conjoint des deux fonctions) et, en second lieu, pour offrir *in fine* la possibilité d'une exploitation des données sur différents supports (site web, cédérom, publication imprimée, etc.).

La décision de proposer un dictionnaire wolof sous format électronique et intégrant une dimension bilingue découle du constat que les dictionnaires monolingues ou bilingues pour cette langue sont non seulement peu nombreux, mais aussi et surtout que l'accès à ces trésors par la population wolophone est handicapé par divers obstacles en termes de publics visés, de coût et de disponibilité. En raison de son format électronique, ce dictionnaire,pourra être librement étendu, exploitable sur divers médias et réutilisé par la suite par d'autres équipes. Quant à la dimension bilingue du projet, elle se manifeste au travers de l'association, pour chaque entrée lexicale wolof, d'une proposition d'équivalent

en langue française (accompagnée d'indications supplémentaires évoquées plus loin).

De par sa durée et les ressources mises en oeuvre, le projet ne prétend pas à l'exhaustivité : il entend plutôt fournir un modèle conceptuel et un canevas technique simples, ouverts et facilement réutilisables pour la production de bases de données et de dictionnaires généraux ou spécialisés tant en wolof que dans d'autres langues qui n'ont guère pu profiter jusqu'à présent des avancées de l'informatique. Remercions enfin Ndeye Fatou Thiaw, Elhadj Diéye, Dame Ndao et Noël Biagui, doctorants à l'UCAD, pour leur regard critique et leur contribution essentielle à toutes les étapes du projet depuis son lancement.

2. CONTEXTE DU PROJET

Comme d'autres langues africaines, le wolof n'a guère bénéficié jusqu'à présent des avantages découlant des avancées de l'informatique depuis la fin des années 1990 en matière d'universalisation du traitement (Unicode) et d'échange (XML) des données textuelles. Or cette langue est non seulement la première langue vernaculaire du Sénégal (et une de celles de la Gambie et de la Mauritanie) mais est aussi une importante langue véhiculaire entre Sénégalais (du moins en termes de *corpus* par contraste avec le français qui occupe, lui, la première place en termes de *status* selon (R. Chaudenson, 1991). Le français étant langue officielle et d'enseignement au Sénégal (à ce jour, il n'existe pas d'enseignement dans les langues nationales en dehors de classes d'expérimentation au niveau de l'enseignement élémentaire), son apprentissage effectif nécessite le développement d'outils didactiques qui prennent en considération le substrat linguistique des apprenants et permettent « d'offrir une passerelle » vers les langues locales. Dans cette optique, le dictionnaire a été élaboré pour tenter de répondre aux besoins lexicographiques de la didactique du français à des wolophones dans l'éducation primaire. À ce niveau d'enseignement, les enseignants ont en effet besoin d'outils facilitant l'apprentissage de l'orthographe et de la signification des mots en wolof (partie unilingue) d'une part, et d'outils facilitant l'apprentissage du français à partir de la langue maternelle des apprenants d'autre part (partie bilingue). Le dictionnaire intègre ainsi les mots les plus fréquents du wolof tels que recensés dans (Diouf, Calvet et Dia, 1971). Il a également pour vocation de faciliter les échanges interculturels du fait qu'il peut être intégré dans une stratégie de didactique du wolof à des francophones.

La disponibilité des données lexicologiques sous forme électronique, en conformité avec les standards Unicode et XML, est une condition indispensable non seulement pour leur exploitation à long terme par les chercheurs au Sénégal et ailleurs, mais aussi pour leur réutilisation et leur intégration dans des applications d'ingénierie linguistique tel qu'un vérificateur orthographique. La forme de stockage des données et leur mode de diffusion pallie, en outre, les inconvénients du support papier en termes de coût, de disponibilité et de diffusion.

3. CONCEPTION DE LA BASE DE DONNÉES LEXICALE

3.1. Outils mis en œuvre

Le principal outil mis en oeuvre pour les besoins du projet est l'outil de base de données *Toolbox* (version 1.5) de SIL International[i]. Ce gratuiciel, utilisé pour la création et l'entretien de la base de données lexicales et dont une version francisée est en cours

d'élaboration dans le cadre d'une collaboration entre SIL International, le LLACAN et le Centre de recherche Termisti, partenaire de l'action de recherche, a été retenu, entre autres, pour sa capacité éprouvée à gérer Unicode et pour ses possibilités d'exportation au format XML.

D'autres outils sont mis à contribution en amont ou en aval, selon les besoins spécifiques des étapes concernées du projet.

En amont, par exemple, la collecte de données lexicales pour la base a été exécutée à partir de textes en wolof numérisés et de l'exploitation subséquente du corpus obtenu à l'aide du concordancier *WordSmith*[ii]. En aval, vu qu'il est prévu que des données audio complètent le dispositif et que, à terme, un fichier son au format *.mp3* soit éventuellement associé à chaque entrée et à chaque phrase d'illustration en wolof et mis à disposition sur le Web, le logiciel *Praat*[iii] sera mis à contribution, entre autres, pour la segmentation des fichiers audio enregistrés au format *.wav* et la production des fichiers *mp3*.

3.2. Constitution d'un corpus textuel restreint

Les moyens et la durée du projet étant limités, une première tâche a été de disposer d'un corpus numérisé du wolof contemporain de taille certes réduite mais recouvrant autant que possible plusieurs domaines fonctionnels. La littérature générale ou spécialisée en langue wolof est, en effet, peu abondante et le nombre de documents accessibles sous forme numérique encore plus réduit.

La majorité des 35 textes dont est constitué le corpus exploité jusqu'ici est donc issue du secteur de l'alphabétisation fonctionnelle (éducation à la santé, à la citoyenneté, etc.) auxquels s'ajoutent des échantillons de la littérature romanesque ou poétique (contes, nouvelles, essais) ainsi que des interviews et la transcription de discours politiques.

Les textes collectés qui n'étaient pas encore numérisés l'ont été de manière à disposer d'une base de données textuelle limitée, mais susceptible de fournir des attestations en contexte des entrées lexicales. Chaque texte a fait l'objet d'une description bibliographique complète de manière à pouvoir disposer d'un en-tête conforme à la TEI (*Text Encoding Initiative*). Le traitement des 35 textes à l'aide de *WordSmith* permet de disposer de statistiques en termes de fréquences (dont les hapax) texte par texte et tous textes confondus de sorte qu'il nous est possible d'identifier le(s) texte(s) spécifique(s) associés à un contexte d'attestation dans la base de données lexicale.

Toutefois, du fait des limites évoquées plus haut, le corpus constitué ne compte que quelque 115 000 formes (*tokens*). L'indexation du corpus à l'aide de *WordSmith* a permis d'isoler quelque 14 700 formes uniques. Comme il n'existe pas, à notre connaissance, de lemmatiseur pour la langue wolof, ces 14 700 formes ont été importées dans la base de données *Toolbox* où l'opération de réduction lemmatique est opérée manuellement à mesure que les transcripteurs dépouillent les formes importées pour pouvoir procéder à la description lexicographique conforme au schéma descriptif des données lexicographiques établi pour notre base.

3.3. Enrichissement du corpus lexical

Nous avons comparé les lexèmes provenant du corpus textuel wolof avec les entrées d'un travail scientifique imprimé traitant des 1 500 mots les plus courants du wolof (Diouf, Calvet et Dia, 1971) ainsi qu'avec deux dictionnaires de référence imprimés (Fal et al., 1990 ; Diouf J.-L. 2003) (désignés sous l'appellation de « matériel de référence » ci-après). Ce processus de confrontation autorise, en effet, un enrichissement du corpus

lexical dérivé du corpus textuel au regard de l'objectif de quelque 5 000 entrées au terme des deux ans du projet. Par conséquent, notre base contient, en premier lieu, des entrées tirées du corpus textuel absentes du matériel de référence. En second lieu, elle contient des entrées communes au matériel de référence et à notre corpus, enrichies d'indications absentes du matériel de référence et, en particulier, d'un contexte et d'une source d'attestation (voire d'une note d'usage le cas échéant). En dernier lieu, dans le cas d'une entrée existante dans le matériel de référence mais absente du corpus textuel, l'entrée intégrée dans la base ne dispose certes d'aucun contexte ni source d'attestation provenant de notre corpus mais est enrichie de toutes les indications liées au schéma de données appliqué à toute entrée de la base, dont une définition et une illustration phrastique créées par l'équipe des transcripteurs du projet de telle manière que le corpus lexical de la base est systématiquement constitué de lexies adossées à un contexte d'attestation et/ou à une phrase d'illustration.

En ce qui concerne l'équivalent français de chaque entrée, lorsqu'il y a désaccord sur l'équivalent proposé par les transcripteurs et/ou des personnes ressources consultées, le matériel de référence est utilisé comme outil de comparaison. Des personnes ressources sont également consultées en cas de désaccord entre transcripteurs sur l'équivalent proposé pour une entrée absente du matériel de référence.

Pour ce qui concerne la terminologie des parties du discours pour le wolof et la définition du wolof fondamental, nous nous sommes appuyés sur des travaux existants du Centre de linguistique appliquée de Dakar et de l'Institut des langues nationales de Nouakchott (Mbodj et Dioulo, 1998 ; Diouf, Calvet et Dia, 1971).

Enfin, pour ce qui concerne la variance orthographique résultant de l'application de règles de transcription différentes, nous nous sommes fondés sur le décret n° 2005-992 relatif à l'orthographe et à la séparation des mots en wolof du 21 octobre 2005 de l'État sénégalais. Relevons que cette précision est d'importance : les dictionnaires imprimés ne précisent pas systématiquement les règles ou usages suivis pour la transcription et, même quand c'est le cas, ne la respectent pas forcément de sorte que d'un dictionnaire à l'autre, la graphie et l'ordre alphabétique des lexies peut varier. Dans notre cas, si le corpus textuel n'a pas été normalisé, le corpus lexical l'est, lui, en fonction du prescrit décrétal.

3.4. Structuration des données lexicales (champs de la base)

Le modèle de données retenu pour la base de données lexicales est orienté par le fait que, le wolof étant considéré comme la langue de référence, l'essentiel de la description lexicographique concerne cette langue. Il est donc prévu que chaque lexème sera accompagné d'informations élémentaires mais pertinentes au regard des objectifs du projet : prononciation (au format mp3), information grammaticale, définition, note d'usage, contexte d'attestation, source(s), synonyme(s), homonyme(s), etc.

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. L'accès aux différents sens en wolof (et à leur équivalent en français) est toujours possible dans la mesure où les entrées sont reliées par le biais d'un champ nommé « homonyme ». Par exemple, le verbe wolof *muus* ayant deux significations (en français, « être rusé » ou « être desséché »), il fera l'objet de deux entrées distinctes. Ce choix a été guidé par la volonté de limiter à un le nombre de significations et d'équivalents d'une

entrée. Dans la même optique, vu que, en wolof, bon nombre de termes appartiennent fréquemment à deux catégories grammaticales différentes et ont de ce fait deux « sens » différents, ils feront l'objet de deux entrées distinctes ; ainsi, la lexie *lekk* étant à la fois un nom (« nourriture ») et un verbe (« manger »), elle fait l'objet de deux entrées dans la base, qui, dans ce cas aussi, seront reliées par le biais du champ « homonyme ».

Le schéma descriptif des entrées repose sur une hiérarchisation en trois niveaux des données (l'outil *Toolbox* permet de définir des relations de subordination entre les descripteurs). 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 capture d'écran ci-dessous (figure 1) donne un aperçu d'une entrée et des champs qui y sont associés dans *Toolbox*.

Lexème wolof	askan
Transcription phonétique	ɛskɛn
Fichier son du lexème wolof	
Catégorie grammaticale du lexème wolof	turu bokkaale
Classe nominale du lexéme wolof	w-
Source du lexème wolof	
Définition du lexème wolof	Mbooleem ñi bokk dëkkandoo
Source de la définition du lexème wolof	
Contexte d'attestation du lexème wolof	Texte juridique
Source du contexte d'attestation du lexème wolof	Déclaration universelle des droits de l'homme (http://www.unhchr.ch/udhr/lang/wol.htm)
Note d'usage du lexème wolof	
Variante du lexème wolof	
Synonyme du lexème wolof	
Homonyme du lexème wolof	askan
Homonyme du lexème wolof	askan
Expression dérivée du lexème wolof	
Lexème source de l'expression dérivée	
*	CC
Traduction francaise du lexème wolof	population
Catégorie grammaticale de la traduction francaise	nom
Phrase d'illustration du lexème wolof	Njaboot nekk na meññeef gu am solo ci askan wi.
Fichier son de la phrase d'illustration	
Traduction francaise de la phrase d'illustration	La progéniture constitue une ressource importante pour la population.
Statut de la fiche	ok
Commentaire	
Auteur du statut de la fiche	AMD

Figure 1 : exemple de fiche lexicale de la base de données Toolbox

On peut y voir que le degré de finesse de l'information est limité aux usages projetés par le projet, qu'il s'agisse de l'usage ultérieur du contenu à des fins de recherche par des spécialistes ou de l'usage des données au titre de dictionnaire. De même, toujours en conformité avec la visée du projet, la richesse des informations lexicales est de loin plus importante en wolof qu'en français, mais ne réduit pas pour autant les données lexicales en relation avec le français à la proposition d'un équivalent à la lexie wolof puisque s'y ajoutent la catégorie grammaticale en français ainsi qu'une traduction de la phrase d'illustration wolof afin, d'une part, de situer la lexie en contexte et, d'autre part, d'offrir un corpus de phrases d'illustration bilingues.

Notons enfin que le modèle comporte cinq champs d'administration de la base de données qui permettent de suivre l'état d'achèvement de chaque fiche, l'identité du gestionnaire de la fiche, les éventuels commentaires sur le contenu de la fiche et enfin la date de dernière modification de la fiche (un champ dont la valeur est gérée de manière automatique par *Toolbox*).

Du point de vue de la méthode de travail, les transcripteurs qui ont préparé le corpus textuel et les autres sources d'entrées lexicales de la base sont aussi les personnes qui complètent chaque fiche. Aucun n'est chargé de compléter la totalité des fiches mais chacun est chargé, selon ses compétences spécifiques, de compléter des champs déterminés par lots alphabétiques de fiches. Le travail des transcripteurs est validé et coordonné par une personne et des réunions à intervalles réguliers permettent à l'équipe du projet de décider d'orientations communes dans la résolution des problèmes rencontrés en cours de rédaction des fiches. Ainsi, alors qu'à l'origine, le modèle comprenait un champ « Auteur » désignant l'auteur d'une fiche complète, celui-ci a été abandonné au profit d'une gestion de la rotation des lots de fiches entre transcripteurs ainsi qu'entre eux et le coordinateur pour l'avancement du projet. Cette gestion présente le double avantage d'un « remplissage » de certains champs selon un fil logique prédéterminé (p. ex. le champ « définition wolof » est toujours complété avant le champ « traduction française ») et la possibilité de retours plus fréquents sur le contenu de champs déjà complétés.

4. MISE A DISPOSITION DES DONNÉES

Au terme du projet, la disponibilité et la diffusion des données auprès des publics visés – chercheurs, linguistes et population wolophone – seront assurées au travers d'une interface web déclinée dans les deux langues du projet, wolof et français, afin d'en renforcer l'accessibilité. En ce qui concerne l'infrastructure technique, les données seront hébergées sur un serveur d'une des institutions partenaires.

Pour se conformer à la double visée du projet, l'interface web proposera le choix entre deux modes d'accès aux données, à savoir un accès aux données brutes aux fins de la recherche linguistique et un accès aux données prétraitées aux fins de la consultation du dictionnaire. La principale différence entre l'un et l'autre modes d'accès est que, dans le premier cas (usage de type « recherche »), le visiteur aura la possibilité de récupérer les données « brutes » ou intégrales exportées depuis *Toolbox* dans divers formats pour une réutilisation à des fins d'ingénierie tandis que, dans le second (usage de type « dictionnaire »), le visiteur aura accès aux données dans un format prédéfini (HTML) pour consultation et, au besoin, pour un transfert dans ce format vers un support de diffusion hors ligne, que ce dernier soit électronique ou imprimé.

En ce qui concerne les données brutes mises à la disposition des chercheurs, les données exportées de *Toolbox* seront librement disponibles pour téléchargement dans un fichier au

format XML produit par *Toolbox* ainsi que dans un fichier au format CSV. Pour faciliter la réutilisation du matériel lexical exporté dans des vocabulaires XML standards tels qu'OLIF, des feuilles de style seront fournies pour l'exécution des conversions nécessaires. Par ailleurs, si les auteurs de textes utilisés pour la constitution du corpus mais ne relevant pas du domaine public nous y autorisent, les chercheurs disposeront aussi de la faculté de télécharger les fichiers texte utilisés pour l'alimentation de la base de données lexicale. La seule restriction au libre téléchargement du matériel du projet concernera les fichiers son (exploitables à des fins d'analyse acoustique), auxquels l'accès se fera sur demande auprès de l'institution chef de file du projet .

Pour ce qui concerne l'usage à des fins de consultation du dictionnaire, l'utilisateur aura la possibilité de consulter le dictionnaire en ligne via une barre d'accès alphabétique classique tant pour le wolof que pour les combinaisons wolof-français / français-wolof. L'utilisateur aura aussi la capacité de définir l'étendue des informations souhaitées, du jeu d'informations réduit à sa plus simple expression à un jeu d'informations complet. Pour la diffusion des données, le visiteur disposera du fichier du dictionnaire complet (pour le wolof d'une part et pour la combinaison wolof-français et français-wolof d'autre part) ainsi que d'un fichier pour chaque lettre de l'alphabet dans les deux langues. Ces fichiers seront fournis au format HTML avec une feuille de style adaptée pour l'impression des données. Enfin, une documentation simplifiée visant les internautes non avertis sera rédigée et mise en ligne afin de documenter la consultation des données hors ligne sur support électronique.

Vu que la totalité du matériel brut est mis à disposition, rien n'interdit à d'autres équipes de mettre au point des interfaces d'interrogation beaucoup plus fines ou spécifiquement adaptées à des besoins lexicaux particuliers.

5. CONCLUSIONS

Sur le plan de la linguistique de corpus, le projet n'a certes pas pour vocation d'innover sur en matière de recherche ou de méthode, mais bien d'offrir une première application faisant appel dans une modeste mesure aux apports de la linguistique de corpus pour une langue qui, pour de multiples raisons, n'a guère pu profiter jusqu'à présent des applications de l'ingénierie linguistique en général et de la linguistique de corpus en particulier. Nous le faisons au départ d'un matériel limité et d'une application exemplative qui, pour modestes qu'ils soient et indépendamment des limites exposées dans cet article, n'en sont pas moins concrets, exploitables et réutilisables.

Outre l'application de type dictionnaire que les utilisateurs pourront consulter en ligne et hors ligne à leurs propres fins pour des usages aussi divers que l'éducation de base et l'enseignement en général, la rédaction, la traduction et toute autre activité en relation avec l'apprentissage ou la pratique écrite du wolof, le projet fournit aux chercheurs un premier corpus lexical wolof informatisé qu'ils ont toute liberté d'enrichir et d'augmenter par la constitution et l'exploitation de corpus textuels plus larges ou plus spécialisés afin de satisfaire, par exemple, le besoin de lexiques bilingues spécifiques dans des domaines tels que la médecine ou l'agriculture, ou d'intégrer le contenu de dictionnaires aujourd'hui uniquement disponibles sur papier qui, quand ils ne sont pas épuisés, sont publiés dans les pays du Nord et, le plus souvent, sont trop coûteux et absents des rayons des librairies du Sud.

Indépendamment des divers usages que des chercheurs pourraient faire du matériel, le projet intègre, de par sa conception même, la possibilité d'être étendu à des applications qui débordent le champ lexical et qui sont couramment utilisées aujourd'hui dans d'autres langues en environnement d'apprentissage ou d'écriture sur ordinateur tels les exerciseurs, les vérificateurs orthographiques ou syntaxiques, etc.

Enfin, nous terminerons en soulignant le fait que la démarche et l'application décrites pour le wolof dans cet article peuvent être utilement réutilisées par les spécialistes d'autres langues qui, à l'instar du wolof, figurent parmi les « parents pauvres » de l'informatisation du traitement des langues et des outils mis au point dans ce cadre. C'est en créant de premiers corpus informatisés – aussi imparfaits soient-ils – dans des formats ouverts et standards que l'on offrira à ces langues la possibilité d'exploiter à leur tour les outils et méthodes élaborés par la linguistique de corpus.

RÉFÉRENCES

Chaudenson R. 1991. *La francophonie : représentations, réalités, perspectives*. Aix-en-Provence : Institut d'Etudes créoles et francophones

Diop A., Calvet M., Dia O. B. K. 1971. *Les cent et les quinze cents mots les plus fréquents de la langue wolof*. Dakar : Centre de linguistique appliquée de Dakar (CLAD).

Diouf J.-L. 2003 *Dictionnaire wolof-français et français-wolof*. Paris : Khartala

R. Chaudenson, 1991

Fal A., Santos R., Doneux J. L. 1990. *Dictionnaire wolof-français suivi d'un index français-wolof*. Paris : Khartala

Mbodj C. & Diolo A. 1998. *Terminologie linguistique et grammaticale wolof = Turalinu làmmiñal róofoo-gi-baat ci wolof*. Dakar / Nouakchott : Centre de linguistique appliquée de Dakar (CLAD) / Institut des langues nationales de Nouakchott (ILN)

[i]Le logiciel Toolbox de SIL International peut être téléchargé librement à l'adresse http://www.sil.org/computing/toolbox/.
[ii]WordSmith Tools, logiciel conçu par Mike Scott de l'Université de Liverpool, est commercialisé par Oxford University Press.
[iii]Praat est un gratuiciel conçu par Paul Boersma et David Weenink de l'Institut des sciences de la phonétique de l'Université d'Amsterdam, téléchargeable à l'adresse http://www.fon.hum.uva.nl/praat/.