Des correcteurs orthographiques pour les langues africaines

Chantal Enguehard, Chérif Mbodj

DES CORRECTEURS ORTHOGRAPHIQUES POUR LES LANGUES AFRICAINES

Chantal ENGUEHARD
Laboratoire d'Informatique de Nantes-Atlantique – Nantes- France
enguehard@lina.univ-nantes.fr
Chérif MBODJ
Centre de Linguistique appliquée de Dakar
Université Cheikh Anta Diop – Dakar-Fann-Sénégal
chembodj@refer.sn

Résumé

Nous rappelons brièvement l'histoire récente des grandes langues africaines, ainsi que les contraintes techniques et surtout économiques qui freinent leur expression dans la sphère de l'information écrite et électronique. Cette situation a des conséquences néfastes sur le développement des pays concernés, en particulier une grande partie de la population est analphabète.

Un environnement logiciel adapté, s'appuyant sur des connaissances linguistiques mémorisées dans un lexique électronique, pourrait répondre en partie aux besoins spécifiques de ces langues. Nous montrons comment un correcteur orthographique s'appuyant sur ce lexique peut participer à la diffusion de connaissances linguistiques, et comment, symétriquement, des outils logiciels doivent aider les linguistes à capitaliser des connaissances dans ce même lexique.

Nous examinons la mise en pratique de ce programme avec le développement effectif d'un correcteur orthographique, la spécification du logiciel de constitution de connaissances linguistiques et la constitution initiale de quelques lexiques.

Mots clefs

Correcteur orthographique ; langues africaines ; outils électroniques pour les langues africaines ; lexique électronique.

Enguehard, C., Mbodj, C., "des correcteurs orthographiques pour les langues africaines". BULAG n° 29: La correction automatique : bilan et perspectives, pp.51-68, 2004.

Abstract

The recent history of the west-African languages and the lack of technical, economical development of the involved countries have considerably hindered their electronic development, with the consequences of the massive analphabetism of the population. An adjusted software environment could help for the compilation and distribution of linguistic knowledge. We plan the development of such a software and detail the first steps of this project.

Key-words

Spelling correction for african languages ; softwares for african languages, african languages, electronic lexicon.

Introduction

Les langues africaines sont peu présentes sur Internet, pourtant certaines sont parlées par une importante population, même si ce n'est pas toujours leur langue maternelle. Ainsi, le bambara au Mali, le wolof au Sénégal et le swahili en Afrique de l'Est sont des exemples de grandes langues de communication (elles sont dites véhiculaires car elles permettent l'intercompréhension entre des personnes de langues maternelles différentes) [Calvet 1981]. Malgré leur importance au sein du continent africain, il apparaît que ces langues sont globalement peu informatisées [Diki-Kidiri 2003]. Plusieurs facteurs (techniques, économiques, sociaux) expliquent cette désaffection mais le poids de l'histoire ne peut être passé sous silence si l'on souhaite comprendre la situation.

1. Situation

1.1. Contexte historique et social

Les colonisateurs des pays africains ont eu des attitudes différentes en ce qui concerne les langues locales. Dans l'ex-Congo belge (aujourd'hui RDC[1]), certaines langues africaines étaient enseignées alors qu'au même moment elles n'avaient aucun droit de cité dans les colonies françaises. Il était même formellement interdit aux élèves d'utiliser leurs langues maternelles, y compris dans la cour de récréation, lorsqu'ils étaient entre eux, sous peine de subir des sévices corporels. Ces comportements ont certainement eu des effets

1 République Démocratique du Congo.

psychologiques très négatifs dans le développement de la personnalité de ces jeunes créatures humaines (qui ont notamment appris que leur langue n'est pas digne de l'école).

La scolarisation dans les langues européennes n'a pas été un succès, et il faut bien constater qu'elle a contribué à aggraver le sous-développement de l'Afrique dans les secteurs économiques et sociaux en maintenant une grande partie de la population dans l'analphabétisme. Cette période est aussi marquée par l'apparition de nombreuses monographies sur les langues africaines, œuvres de missionnaires ou d'administrateurs coloniaux, c'est-à-dire de personnes certes de bonne volonté mais sans aucune qualification en linguistique[2]. Il faut souligner que ces travaux d'amateurs ont fini par imposer l'utilisation de l'alphabet latin alors qu'il existait déjà des alphabets autochtones bien adaptés aux langues locales. Citons l'écriture des Bamum que le roi Njoya crée et fait prospérer à la fin du dix-neuvième siècle au Cameroun, ou les syllabaires mandé qui apparaissent entre 1833 et 1930 [Dalby 1986].

Les intellectuels africains ont identifié le péril que représente cette situation de négation des langues et cultures africaines car ils connaissent l'importance primordiale des langues[3]. Ils ont également compris très tôt que l'alphabétisation de la population est un facteur essentiel pour développer un pays, et qu'il est nécessaire d'alphabétiser la population dans une langue qu'elle comprend, et non plus dans une langue totalement étrangère comme celle du colonisateur[4]. Il faut donc que le travail de description scientifique des langues africaines soit mené par des linguistes de métier afin de mettre au point des systèmes de transcription susceptibles de fixer les langues locales et d'aider à leur introduction dans le système

2 Ainsi, le même son [u] est transcrit 'u' par les colonisateurs anglophones, et 'ou' par les francophones.

3 Une langue qui disparaît s'accompagne de la disparition de la culture partagée par la population dont c'était la langue maternelle [Calvet 1987].

4 Rappelons, à cet effet, que dès 1817, Jean Dard (*cf.* Gaucher 1968), instituteur français en poste à l'école mutuelle de Saint-Louis du Sénégal dont il était le directeur, avait essayé d'utiliser le wolof comme langue d'enseignement, après avoir constaté « l'échec d'un enseignement unilingue français amenant les élèves à lire et à écrire le français sans le comprendre » (*cf.* Dumont 1983) . Malheureusement, la commission scolaire que le gouverneur Dubelin envoya, le 25 mars 1829, à l'école mutuelle de Saint-Louis condamna et mit fin à cette innovation pédagogique qui desservait la politique assimilationniste de la France d'alors.

éducatif.

Nous ne pouvons évoquer les circonstances de ce nécessaire effort linguistique pour chacune des langues de chacun des pays d'Afrique de l'Ouest et nous restreignons pour cet article au cas du Sénégal.

Le 10 septembre 1937, lors de sa conférence à la Chambre de commerce de Dakar, le futur Président de la République, Léopold Sédar Senghor, préconisa l'enseignement des langues maternelles. Après l'accession de son pays à la souveraineté internationale, il donna des instructions afin que dans, une première étape, six des langues africaines parlées au Sénégal (wolof, sereer, jóola, manding (malinke) et soninke) soient élevées au statut de langues nationales et fassent l'objet de décrets réglementant leur transcription et le découpage des mots en leur sein.

Malheureusement, il convient de remarquer qu'aujourd'hui les systèmes orthographiques officiels ne sont pas harmonisés d'un pays à un autre pour les mêmes langues, voire à l'intérieur d'un même Etat ; ce qui a conduit, naturellement, à l'existence d'usages différents à l'intérieur d'une même langue. On pourrait illustrer cette situation par plusieurs exemples empruntés au wolof[5] : l'énoncé [nja :y][6] est transcrit *ndiaye* au Sénégal, et *ɲjie* en Gambie[7]. Une telle situation a été évitée entre la Mauritanie et le Sénégal, dont les systèmes de transcription étaient respectivement fondés sur des principes d'ordre morphophonologique ou phonologique grâce à un projet de l'ACCT[8] (actuelle AIF[9]) qui a permis, dans les années 90, à des experts sénégalais et mauritaniens de se réunir, à Dakar comme à Nouakchott, et d'harmoniser le système de transcription du wolof dans les deux pays concernés.

A ce problème d'harmonisation il s'ajoute que, faute d'avoir été rendus obligatoires par des mesures législatives coercitives, les systèmes de transcription officiels ne sont pas toujours respectés et

5 Le wolof est une langue parlée à la fois au Sénégal et en Mauritanie, colonisés par les Français, et en Gambie, colonisée par les Anglais.

6 Nom de famille wolof.

7 Ainsi, une unique langue est transcrite de deux manières différentes suivant le pays où elle est parlée ce qui complique de manière absurde la communication écrite au sein d'une même communauté linguistique [Mbodj 2002].

8 Agence de Coopération culturelle et technique.

9 Agence intergouvernementale de la Francophonie.

les décrets sont généralement laissés pour compte.

Mais que constate-t-on, en outre, à l'heure des TIC[10] ? Que non seulement les systèmes de transcription ne sont pas harmonisés, mais également que la représentation électronique usuelle des caractères spéciaux utilisés dans les alphabets n'obéit à aucun standard. Ces écueils majeurs ne permettent pas à ces langues de participer pleinement au monde « du donner et du recevoir » que constitue le cyberespace.

Or, les grandes langues africaines, autant dire un pan important de l'Humanité, aspirent à accéder à cet espace globalisé de communication et d'échange qu'est le cyberespace. D'où l'importance que revêt la réalisation de correcteurs orthographiques pour les langues africaines dans la promotion des décisions linguistiques d'harmonisation des systèmes d'écriture.

1.2. Codage des caractères spéciaux

Les langues africaines sont souvent transcrites à l'aide d'alphabets latins comprenant en sus quelques caractères spéciaux (exemples : ɔ, ɛ, ə, ŋ). Ces caractères ne font pas partie des tables d'encodage électronique courantes (ascii sur 7 bits, ou ascii étendu sur 8 bits). Pour pallier cette absence, ils ont été dessinés à la place de caractères existants, mais dont on estime ne pas avoir besoin pour écrire une langue particulière. Ainsi, nous voyons sur la table 1 que dans la police Bambara Arial la lettre 'q', absente de l'alphabet bambara, a été redessinée afin d'obtenir l'affichage de la lettre 'ɛ' (epsilon) à sa place.

10 Technologies de l'Information et de la Communication.

	Caractère initial			
Caractère affiché	Alphafrica	Arial Bambara	Bambara Arial	Times New bambara
ɔ	µ	<	ù	<
Ɔ	Ó	>	%	>
ɛ	*f*	&, ²	q	&, ²
Ɛ	·	^, µ	Q	^, µ
ɲ	≈	$	x	$
Ɲ	⁄	%, §	X	%, §
ŋ	¬	#	v	#
Ŋ	|	@	V	@

Table 1 : représentation des caractères spéciaux du bambara dans 4 polices de caractères

Cette solution s'est largement développée puisqu'elle a longtemps représenté la seule possibilité pour écrire avec les outils électroniques disponibles, mais elle présente de nombreux inconvénients. Tout d'abord, elle complique l'échange de fichiers de textes à cause de l'absence de consensus dans la mise au point de telles polices de caractères alternatives. Ainsi, un texte écrit avec un ordinateur disposant de la police A et affiché sur un autre ordinateur, avec la police B, sera illisible (*cf.* Table 2). Tout envoi de fichiers doit donc être accompagné des polices de caractères utilisées par ces fichiers[11].

Alphafrica	Cɛnimusoya ye jɛŋɔgɔnya wale ye min kɔnɔ cɛni muso bɛ jɛ fo den bɛ se ka bɔ a kɔnɔ.
Bambara Arial	C*f*nimusoya ye j*f*≈µgµnya wale ye min kµnµ c*f*ni muso b*f* j*f* fo den b*f* se ka bµ a kµnµ.

Table 2 : affichage d'un même texte avec les polices de caractères Alphafrica et Bambara Arial

11 Il est fréquent qu'un fichier utilise plusieurs polices de caractères puisque les caractères redessinés ne sont plus affichables. Ainsi, un texte mathématique écrit en bambara, à l'aide de la police de caractères Alphafrica, et dans lequel est utilisé 'µ' (lettre grecque mu), devra obligatoirement utiliser une autre police de caractères pour afficher 'µ' puisque ce caractère est redessiné ɔ (o ouvert) dans la police Alphafrica.

En second lieu, cette représentation interdit tout traitement automatique des langues puisque le code de chaque caractère n'est pas fixé, le même code pouvant être utilisé par plusieurs caractères (sur la table 1, par exemple, nous voyons que 'ε' (epsilon) occupe la place du '&' dans la police Arial Bambara et celle du 'q' dans la police Bambara Arial).

En 1992, a émergé le standard Unicode, fruit d'une concertation entre industriels membres du Consortium Unicode et les représentants de l'Organisation internationale de normalisation (ISO) [Andries 2002]. Il s'agit d'un système de codage qui peut être étendu sur 2, 3 ou 4 octets. Il permet de représenter plus d'un million de caractères différents, c'est-à-dire tous les caractères de toutes les langues.

La mise au point et la diffusion de ce standard constituent un progrès considérable puisqu'il autorise toutes les langues à franchir la première étape de l'informatisation d'une langue : le stockage des documents sous une forme électronique qui permette leur traitement analytique [Chanard 2001].

De nombreux outils adaptés à unicode commencent à apparaître.

1.3. Outils électroniques

Claviers

La plupart des caractères spéciaux ne figurent pas sur les claviers couramment distribués. Il est possible de les saisir en utilisant leur code, mais cette solution manque évidemment d'ergonomie (il faut se souvenir des codes, appuyer sur plusieurs touches pour obtenir un caractère). Dans le cadre d'une action de recherche en réseau de l'AUF réunissant l'Université de Nouakchott, l'Université de Dakar et l'ISTI, des claviers virtuels Unicode ont été développés en balante, bambara, pulaar, sereer et wolof[12]. Ces claviers permettent d'obtenir directement les caractères spéciaux requis par la frappe de touches du clavier ; le code généré est le code unicode du caractère.

Editeurs de textes

Les éditeurs de textes couramment disponibles (comme Word ou

12 http://www.termisti.refer.org/ltt/ltt.htm

Open Office) sont réalisés dans des langues de statut international (anglais, français, espagnol, etc.), il est évidemment possible d'utiliser de tels éditeurs pour écrire d'autres langues malgré quelques difficultés.

Tout d'abord, il faut maîtriser la langue dans laquelle est rédigée son interface. L'utilisateur est obligé de fonctionner en mode bilingue, ce qui n'est peut-être pas sans conséquence sur son fonctionnement cognitif, l'une des langues pouvant influencer les mots et structures syntaxiques choisis pour la rédaction dans l'autre langue. Ensuite, les fonctionnalités linguistiques complémentaires, comme la correction automatique de l'orthographe, sont inexistantes pour certaines langues (même imparfait, il est évident qu'un correcteur orthographique représente un soutien appréciable pour améliorer la qualité d'un texte). Enfin, les textes produits sont codés en ascii et affichés à l'aide de polices de caractères éventuellement redessinées.

Les progrès récents permettent d'envisager le développement d'éditeurs de textes adaptés aux langues africaines et produisant des textes au format unicode.

1.4. Ressources linguistiques rares

Ecrire dans une langue africaine, avec ou sans ordinateur, reste un exercice difficile. Dans la grande majorité des cas, il n'existe pas de ressources linguistiques imprimées permettant de trancher les questions sémantiques ou syntaxiques. Ainsi, la plupart des langues ne bénéficient d'aucun dictionnaire monolingue[13], ce qui est une situation paradoxale puisque qu'elles sont souvent dotées de plusieurs dictionnaires bilingues.

Bien que cruciales, les étapes de production et de distribution de ressources linguistiques paraissent hors de portée. La production d'un dictionnaire monolingue représente un travail titanesque, or les personnes qualifiées sont rares en Afrique et généralement déjà mobilisées sur des tâches également importantes comme la production de manuels d'éducation ou de santé. De plus, la diffusion des ouvrages produits est confidentielle car ils restent onéreux, et la population est globalement peu alphabétisée !

13 Nous pouvons citer un dictionnaire monolingue zarma : Isufi Alzuma Umaru, *Kaamuusu Kayna*, éd. Alpha, 1996.

Face à ce décourageant constat, l'utilisation systématique des ordinateurs lors de la production de textes (qu'il s'agisse de saisie ou même d'élaboration directe) représente une opportunité à saisir car elle offre la possibilité de diffuser et de recueillir des connaissances linguistiques via un éditeur de texte adapté, et plus particulièrement grâce à un correcteur orthographique adapté.

2. Spécification d'un correcteur orthographique adapté

2.1. Bref état de l'art

Les correcteurs orthographiques constituent un axe de recherche depuis les années 1960 [Kukich 1992]. Ils sont maintenant couramment utilisés par le grand public car les éditeurs de textes courants en intègrent souvent un, et qu'ils apportent un confort non négligeable lors de la rédaction de textes. Ces correcteurs fonctionnent selon un mode interactif dans lequel intervient l'utilisateur, contrairement aux correcteurs orthographiques complètement automatiques comme dans le domaine de la reconnaissance optique de caractères (et dont nous ne nous préoccupons pas ici).

Un correcteur orthographique interactif fonctionne en suivant plusieurs étapes :

- détection des erreurs ;
- sélection des corrections possibles ;
- ordonnancement des corrections possibles et proposition à l'utilisateur ;
- correction effective du texte respectant le choix de l'utilisateur.

● La détection des erreurs s'effectue souvent en considérant un à un les mots du texte à corriger, de manière isolée. Chacun des mots du texte est comparé aux mots du lexique (qui contient les mots de la langue, ainsi que leurs flexions). Tout mot non trouvé dans le lexique est considéré comme erroné. Cette technique est très simple à mettre en œuvre mais présente l'inconvénient de ne pas détecter les erreurs transformant un mot en un autre mot présent dans le lexique comme dans la phrase « le livre est sue la table ». Le mot « sur » (préposition), a été transformé en « sue » (verbe suer), ce qui est manifestement erroné. Le taux de telles erreurs non détectées

augmente avec l'accroissement de la taille du lexique, car plus celui-ci contient de mots, plus il est possible qu'une erreur transforme un mot en un autre mot du lexique. L'augmentation de la taille du lexique contribue donc, paradoxalement, à dégrader les performances du correcteur orthographique. Seule la prise en compte du contexte d'apparition des mots peut aider à éviter cet écueil majeur. Les premières expériences dans ce sens étaient fondées sur le calcul de trigrammes (sur les mots). Cette approche théoriquement valide présente l'inconvénient majeur de nécessiter un énorme corpus d'entraînement qu'il n'est pas toujours possible de constituer [Mays 1991]. Les plus récents travaux s'inscrivent dans le domaine de la cohésion lexicale. Tester simplement les chaînes lexicales sur une phrase aboutit à ce que le système détecte des erreurs qui, pour les neuf dixièmes, n'en sont pas [Hirst 1998]. Un nouvel algorithme exploitant les relations sémantiques diverses que peuvent entretenir les mots (synonymie, méronymie, fréquence de cooccurrences élevée, etc.), et étendant la notion de voisinage, autrefois restreinte à la phrase à un ou plusieurs paragraphes, semble capable de bien meilleures performances [Hirst 2003].

- Quand une erreur est détectée, le correcteur sélectionne une série de mots susceptibles d'être la version correcte de la chaîne à corriger. Ces mots sont choisis selon différentes techniques (calcul de la distance minimale d'édition, clé de similarité, ou encore mesure de la distance phonologique).

- L'ordonnancement des chaînes candidates à la correction prend en compte la mesure utilisée lors de l'étape de sélection, ainsi que des mesures statistiques (comme la fréquence d'apparition des mots, ou bien le mot le plus fréquemment choisi lors de rencontres préalables avec la même erreur).

- Enfin, une étape interactive permet à l'utilisateur de superviser la correction. Il peut adopter l'une des trois attitudes suivantes :

    - corriger le mot erroné en sélectionnant un des candidats proposés par le correcteur ;
    - modifier le mot erroné ;
    - ne pas corriger ; dans ce dernier cas, il peut rajouter ce mot à son dictionnaire personnel.

Les correcteurs orthographiques rencontrent deux difficultés majeures. Tout d'abord, les concaténations intempestives de mots, ou l'insertion d'un délimiteur (caractère espace, ponctuation) à l'intérieur d'un mot rendent très délicate la sélection de candidats pour la correction. Cette difficulté n'est cependant pas trop gênante dans le cadre d'un fonctionnement interactif car ces erreurs de frappe sont facilement corrigées par l'utilisateur. La mise à jour du lexique constitue un écueil plus important : les langues évoluent assez vite comme le montre le grand nombre d'ajouts et de suppressions de mots lors des révisions annuelles des dictionnaires destinés au grand public, l'utilisation d'un correcteur fondé sur un lexique vieux de plusieurs années révèle que nombre de mots couramment utilisés sont faussement diagnostiqués comme erronés car ce sont des emprunts, des néologismes, ou de nouvelles dérivations de mots.

2.2. Inadéquation des correcteurs orthographiques existants pour les langues africaines

Il existe déjà des correcteurs orthographiques pour certaines langues africaines, mais ils sont généralement très simples : il s'agit d'utiliser des correcteurs orthographiques existants en leur fournissant un lexique correspondant à la langue visée [Van der Veken 2003]. Ces correcteurs orthographiques localisent les erreurs en scrutant les mots du texte de manière isolée et, même s'ils rendent des services appréciables, ils rencontreront fatalement les problèmes précédemment soulignés. Nous pensons qu'un correcteur orthographique adapté aux langues africaines doit prendre en compte les contextes des mots afin de ne pas limiter inévitablement ses performances.

Par ailleurs, il doit posséder des fonctionnalités supplémentaires par rapport aux correcteurs orthographiques habituels afin de prendre en compte le contexte de dénuement linguistique des langues africaines. D'une part, il peut participer à la diffusion de connaissances linguistiques en accompagnant les propositions de corrections d'informations linguistiques supplémentaires, d'autre part, il peut aider à la constitution de ressources linguistiques en recueillant des données destinées à des linguistes.

2.3. Spécification d'un correcteur orthographique adapté

Nous avons choisi de réaliser un correcteur orthographique simple,

compatible avec le standard unicode, et fonctionnant avec des ressources linguistiques limitées et incomplètes. Nous avons défini deux fonctionnalités supplémentaires pour, d'une part, communiquer davantage d'informations linguistiques à l'utilisateur et, d'autre part, encourager la capitalisation de ressources linguistiques.

● Lors de la correction d'un mot détecté comme erroné, le correcteur propose des mots candidats à la correction. Comme les langues africaines sont peu standardisées et présentent de nombreuses variantes dialectales, en particulier phonologiques, il est possible que l'utilisateur n'identifie pas certains mots proposés car ils sont orthographiés d'une manière qui lui est peu familière (mais qui est officielle) ou qu'il n'a jamais rencontrée. La communication d'informations supplémentaires sur les mots (comme leur(s) définition(s), leur catégorie grammaticale, des exemples d'usage, etc.) pourraient l'aider à choisir le mot adéquat et à l'utiliser correctement.

● Un correcteur orthographique rencontre, par définition, de nombreux textes, et est muni d'un lexique qui lui permet d'identifier les mots absents du lexique (mots désignés comme a priori erronés). Son fonctionnement prévoit qu'un utilisateur peut ajouter des mots corrects, mais absents du lexique général, à son lexique personnel. Nous souhaitons exploiter ce processus d'enrichissement afin d'aider les institutions en charge des langues à augmenter le lexique officiel disponible pour une langue. Lors de l'ajout d'un mot au lexique personnel, le correcteur orthographique peut mémoriser ce mot dans un fichier, ainsi que la phrase dans laquelle il apparaît. L'utilisateur est vivement encouragé à transmettre ce fichier à l'institution en charge de la langue. Celle-ci peut utiliser les informations contenues dans ce type de fichiers pour enrichir le lexique de la langue (*cf*. aide à la production de ressources linguistiques).

Ce correcteur orthographique est, dans une certaine mesure, indépendant de la langue puisque nous décrivons les traitements dépendants de la langue (comme le calcul des flexions et dérivations des mots par exemple) dans des modules génériques qui utilisent les informations contenues dans les ressources linguistiques rassemblées dans le lexique électronique de la langue. Pour adapter ce correcteur orthographique à une nouvelle langue, il suffit donc de changer de lexique.

2.4. Le lexique

Le lexique est placé au centre de notre démarche car il est utilisé par le correcteur orthographique comme ressource, qu'il doit être maintenu et enrichi par l'institution en charge de la langue et, qu'en l'absence de ressources linguistiques imprimées, il est souhaitable que l'utilisateur puisse le consulter directement. Donc, il peut contenir des informations qui ne sont pas directement exploitables par un correcteur orthographique, mais qui sont importantes pour un utilisateur humain. De manière symétrique, il contient des informations statistiques qui ne sont pas directement utilisables par un utilisateur.

La diversité des usages d'un tel lexique nous a poussés à spécifier plus précisément les interactions auxquelles il participe et à proposer des aides logicielles adéquates :

- Lors de la correction d'un texte, l'utilisateur manipule un correcteur orthographique qui consulte le lexique.
- L'utilisateur peut souhaiter consulter le lexique hors de toute tâche de correction. Il a besoin dans ce cas d'un logiciel permettant de feuilleter le lexique. La transformation automatique du lexique en un document hypertexte répond à cette préoccupation.
- L'institution en charge de la langue doit enrichir le lexique soit en introduisant de nouveaux mots, soit en complétant les informations sur les mots qui en font déjà partie. Nous avons déjà décrit une voie d'enrichissement possible avec la proposition de nouveaux mots par les utilisateurs du correcteur orthographique. Les mots nouveaux pourraient également être suggérés par l'analyse de corpus. L'institution en charge du lexique doit donc disposer d'un logiciel lui offrant le plus d'aide possible pour la maintenance du lexique. Nous décrivons ce logiciel en détail dans la partie suivante.

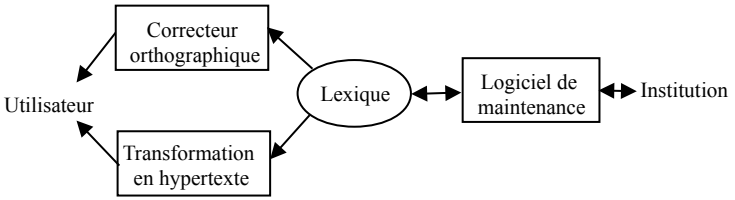

Fig. 1 : Interactions entre le lexique, l'utilisateur et l'institution en charge de la langue

Un lexique est défini comme un ensemble d'items. Chaque item regroupe un ensemble d'informations dont le radical de la forme, la catégorie grammaticale, le mode de flexion, une définition, un ou plusieurs exemples d'usage. D'autres informations peuvent être ajoutées, qui ne seront pas forcément utilisées par le correcteur orthographique (phonétique, synonymes, etc.).

Exemple d'item (wolof) :

radical de la forme :	aay
phonétique :	[a :y]
catégorie grammaticale :	v.i.
mode de flexion :	2
définition :	Être mauvais, être mal
exemples d'usage :	Lu aay ci li ma wax. (Qu'est-ce qu'il y a de mal dans ce que j'ai dit ?)
synonyme :	bon

Dans notre approche, le lexique n'est pas statique mais est perpétuellement « en cours d'élaboration » : plusieurs personnes peuvent intervenir à différents moments pour l'enrichir, le modifier. Il convient donc de mentionner, pour chaque champ, la source de l'information, l'identité de la personne l'ayant validée, ainsi que la date de validation. Ces méta-informations sont très classiques dans le domaine de la lexicographie et permettront d'avoir un retour d'expérience sur la constitution du lexique.

Le lexique est mémorisé au format XML, ce qui laisse la possibilité de l'adapter à différentes normes (comme la norme : « lexiques pour le TAL »en cours de définition[14]).

14 Dans le cadre de l'action SYNTAXE de l'INRIA, du projet RNTL -Outilex, de

Le lexique apparaît ici comme centralisant l'ensemble des informations sur la langue, et cette tendance devrait s'accentuer avec l'introduction d'expressions lexicales et d'informations syntaxiques plus précises.

2.5. Aide à la production de ressources linguistiques

Nous faisons l'hypothèse que l'enrichissement du lexique au sein de l'institution en charge de la langue peut être facilité par le développement d'un logiciel adéquat. Ce logiciel a plusieurs objectifs :

- intégration des contributions des utilisateurs ;
- enrichissement des items du lexique (catégorie grammaticale, définitions, exemples d'usage, etc.) ;
- intégration de corpus pour calculer des informations statistiques (trigrammes sur les symboles, fréquence d'apparition des mots, etc.), observer les contextes d'usage (concordancier), etc.

Les personnes chargées de la maintenance des ressources électroniques verraient leur tâche facilitée par l'utilisation d'un tel logiciel leur permettant d'observer les mots en contexte (grâce à la prise en compte d'un corpus) et de noter de nouvelles informations (telles la catégorie grammaticale, une définition, etc.) dans des formulaires adaptés.

3. Réalisation

3.1. Recueil de données linguistiques

Nous avons pour projet de développer des correcteurs orthographiques pour plusieurs langues africaines (dans un premier temps : bambara, kanuri, tamajaq, fulfulde (peul), wolof, hausa et zarma). Pour chacune de ces langues, nous avons recueilli des ressources textuelles[15] auprès d'institutions, de journalistes et de maisons d'édition afin d'initialiser le lexique électronique de chacune des langues. Des linguistes spécialistes de ces langues ont

l'action Normalangue et du groupe de travail AFNOR : « lexiques pour le TAL ».

15 Il faut remarquer que les textes recueillis dans 7 langues (*cf.* Table3) utilisent 25 polices de caractères différentes ! Leur exploitation nécessite donc de les transformer en conformité avec le standard unicode.

vérifié que ces textes sont écrits conformément aux règles de transcription et d'orthographe en vigueur afin de ne pas biaiser la qualité de notre correcteur.

Langue	Nombre de mots
bambara	89 684
kanuri	79 336
tamajaq	350 010
fulfulde	24 088
hausa	139 239
zarma	74 398
wolof	en cours d'évaluation

Table 3 : corpus recueillis

Ces ressources textuelles sont complétées par quelques lexiques généralement bilingues (lexique du droit en bambara-français, par exemple), et des dictionnaires également bilingues (bambara-français, hausa-français, wolof-français) ou monolingues (zarma).

Les dictionnaires constituent une ressource particulièrement précieuse car de nombreux liens sémantiques utiles pour une correction orthographique de qualité y sont notés (synonymie, antonymie, analogie)

3.2. Logiciel

Le correcteur orthographique lui-même est en développement à l'Université de Nantes. Il procède par l'observation du texte mot à mot, où chaque mot est comparé aux mots présents dans le lexique (après calcul des formes fléchies). Les candidats proposés à la correction sont les mots du lexique les plus proches de la chaîne erronée au sens de la distance minimale d'édition [Wagner 1974]. La recherche de ces candidats est facilitée par la représentation interne du lexique comme un arbre lexicographique [Oflazer 1996] décoré, les décorations correspondants aux informations utiles lors de la correction (mode de flexion, définition, exemples d'usage, etc.).

Dans une seconde version, le contexte des mots sera pris en compte grâce aux liens sémantiques issus des dictionnaires, et à des mesures effectuées en corpus (fréquences de cooccurrences).

Le correcteur traite des textes sauvegardés au format HTML. Ce mode de fonctionnement présente l'avantage de le rendre compatible avec de nombreux éditeurs de texte puisque les fonctionnalités « sauver au format HTML » et « lire un texte écrit en HTML » sont largement répandues dans les éditeurs courants. Ce format présente l'avantage de mentionner explicitement les polices de caractères utilisées lors de l'élaboration du texte, et donc de rendre possible le recodage d'un texte ascii utilisant une police redessinée en un texte respectant le standard unicode.

Le développement de la première version du correcteur orthographique et du logiciel de maintenance du lexique devrait être achevé en 2004. Les tests qui seront effectués au début de l'année 2005 déboucheront sur la réalisation de la version finale. Celle-ci devrait être disponible, et gratuitement téléchargeable, en 2005.

Conclusion

Nous pouvons déjà prévoir une période de transition jusqu'à un environnement logiciel entièrement compatible avec unicode. Cette période délicate verra cohabiter des outils anciens avec des outils "unicodisés" : un texte entièrement codé selon le standard unicode sera illisible par les anciens outils, les claviers unicode seront inutilisables avec les anciens outils, etc. Il convient donc de développer des passerelles entre ces deux environnements permettant, en particulier, de convertir facilement les anciens documents au nouveau standard, mais également de faire l'opération inverse afin de continuer à bénéficier des outils anciens pendant cette période de transition.

Nous soulignons que certaines innovations décrites, comme la présentation d'informations supplémentaires à l'utilisateur d'un correcteur orthographique, seraient utilisables dans d'autres langues où elles rendraient des services, notamment aux rédacteurs écrivant dans une langue qui n'est pas leur langue maternelle. Les langues africaines jouent donc un rôle stimulant en nous obligeant à faire face à des situation extrêmes. Nous pouvons mettre en parallèle ce transfert d'innovation d'un environnement sociétal à un autre avec l'invention de la télécommande pour télévision inventée en premier lieu pour répondre à un besoin des handicapés, et largement appréciée de l'ensemble de la population.

Références

ANDRIES, P. (2002). " Introduction à Unicode et à l'ISO 10646 ", *Document numérique*, vol.6, n°3-4, pp.51-88.

CALVET, L.-J. (1981). " Les langues véhiculaires ", *P.U.F. coll. Que sais-je ?*.

CALVET, L.-J. (1987). " La guerre des langues et les politiques linguistiques ", *Payot*.

CHANARD, C. et A. POPESCU-BELIS, (2001). " Encodage informatique multilingue : application au contexte du Niger ", *Les cahiers du RIFAL*, n°22, pp.33-45.

DALBY, D. (1986). " L'Afrique et la lettre ", *Centre Culturel Français, Lagos & Fête de la Lettre*, Paris.

DIKI-KIDIRI, M. et E. ATIBAKWA BABOYA E. (2003) " Les langues africaines sur la toile ", *Les cahiers du RIFAL - Le traitement informatique des langues africaines*, n°23, pp.5-32.

DUMONT, P. (1983). " Le français et les langues africaines au Sénégal ", Paris, Acct-Karthala.

HIRST, G. et D. ST-ONGE, (1998). " Lexical chains as representations of context for the detection and correction of malapropisms ", Fellbaum, pp.305-332.

HIRST, G. et A. BUDANITSKY (2003). " Correcting real-word spelling errors by restoring lexical cohesion ", *http://www.cs.toronto.edu/compling/Publications/Abstracts/Papers/Hirst+Budanitsky-2001-abs.html*.

KUKICH K. (1992). " Techniques for automatically correcting words in text ". *Computing Survey*s, 24(4): 377–439.

MAYS, E., DAMERAU, F. J. et R. L. MERCER (1991). " Context based spelling correction ", *Information Processing and*

Management, 27(5), pp.517-522.

MBODJ, C. (2002). " Orthographe commune et législations nationales ", *Writing African – The Harmonisation of Orthographic Conventions in African Languages*, ed. Kwesi Kwaa Prah, pp. 55-64.

OFLAZER, K. (1996). " Error-tolerant Finite-state Recognition with Applications to Morphological Analysis and Spelling Correction ", *Computational Linguistics*, vol. 22, n°1, pp.73-98.

VAN DER VEKEN A. et A. DE SCHRYVER (2003). " Les langues africaines sur la Toile : études des cas haoussa, somali, lingala et isixhosa ", Les cahiers du RIFAL n°23 , pp.33-45.

WAGNER, R.A. et M.J. FISCHER (1974). " The string-to-string correction problem ", *Journal of the Association for Computing Machinery*, vol.21, n°1, p.168-173.