Langues africaines et communication électronique développement de correcteurs orthographiques

Chantal Enguehard, Soumana Soumanak@Yahoo.Com Kané

Langues africaines et communication électronique

développement de correcteurs orthographiques

Chantal Enguehard
Laboratoire d'Informatique de Nantes-Atlantique – Nantes- France
enguehard@lina.univ-nantes.fr

Soumana Kané
Centre National des Ressources de l'Education Non Formelle – Bamako – Mali
soumanak@yahoo.com

Résumé

Les langues d'Afrique de l'Ouest sont peu utilisées pour la communication électronique. Nous examinons différentes causes de cette situation, comme les problèmes liés à la représentation électronique des caractères spéciaux utilisés dans ces langues, l'histoire de la transcription de ces langues, et leur enseignement. Nous préconisons le développement d'outils informatiques adaptés (éditeurs de textes, claviers) afin de favoriser l'écriture de textes. Enfin, nous détaillons les spécifications d'un correcteur orthographique adapté à la problématique exposée. Nous focalisons plus particulièrement sur le cas de la langue bambara (surtout parlée au Mali).

1. Situation

1.1. Communication électronique et diversité linguistique

La communication électronique comprend les échanges de courriers électroniques, l'élaboration de sites Internet, etc. Dans tous les cas, ces activités utilisent un logiciel (éditeur, navigateur, etc.) implanté sur un ordinateur. La communication se déroule via une interface de l'humain vers la machine (clavier, souris), et une interface de la machine vers l'humain (affichage). A chacune de ces composantes se rencontrent des difficultés dues à la présence de caractères spéciaux dans la plupart des langues africaines.

- **Problèmes de codage**

L'échange de messages dans des langues utilisant des caractères différents de l'alphabet latin a été longtemps difficile : ainsi il arrive encore qu'un message en français utilisant quelques signes diacritiques soit mal restitué à sa réception, car les caractères accentués sont remplacés par d'autres caractères. Ce problème s'explique par l'histoire de l'informatique. Les premiers ordinateurs ont été développés par des Américains qui se sont naturellement exprimés en

Enguehard, C., Kané, S., "Langues africaines et communication électronique : développement de correcteurs orthographiques", premières Journées scientifiques communes des réseaux "Langues" de l'AUF, Ouagadougou, 31 mai – 1er juin 2004.

anglais, langue qui n'utilise que les caractères latins. Pour chaque lettre, il a été prévu un code. Comme il était très coûteux de conserver des informations dans des mémoires, le code devait être aussi petit que possible, il a donc été pensé qu'un octet (regroupement de 8 bits) serait amplement suffisant puisqu'il permet de distinguer 256 caractères différents. Au début plusieurs codes ont coexisté, inventés de-ci de-là par les chercheurs. Finalement, la table ASCII[1] s'est imposée. Elle comprend les 26 lettres de l'alphabet latin, quelques lettres avec diacritique, les chiffres, les signes de ponctuation, et quelques codes réservés. Chaque caractère a un code, choisi arbitrairement.

En cas de besoin d'un caractère non présent dans la table ascii, il restait possible de redessiner le glyphe[2] correspondant au code d'un caractère inutile. Malheureusement, dans ce domaine, aucun consensus n'a émergé, le même code a pu être utilisé pour des caractères différents, autrement dit, le même caractère n'est pas représenté par un code unique dans différents documents, ce qui peut poser des problèmes lors de l'échange d'informations[3].

Le standard Unicode a émergé en 1992, 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 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 donc 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].

- Interface machine vers humain

Il s'agit de l'affichage de textes.

La solution habituelle, déjà utilisée pour les caractères avec diacritique, a été adoptée : les glyphes de certains caractères, non utilisés dans la langue, ont été redessinés afin de produire les caractères spéciaux souhaités[4] (les caractères redessinés sont choisis, autant que possible, parmi les caractères présents sur le clavier[5] afin de rendre la saisie plus aisée). Cette solution, très répandue puisqu'elle a longtemps été la seule techniquement réalisable, présente de nombreux inconvénients :

---

[1] http://www.purebasic.com/french/documentation/Reference/ascii.html

[2] Le glyphe est l'image qui correspond au dessin du caractères

[3] Par exemple, le caractère 'œ' (absent de la table ASCII) envoyé dans un mèle s'affiche comme une barre verticale '¦' à sa réception.

[4] La lettre q, par exemple, n'est pas utilisée en bambara.

[5] C'est-à-dire obtenus par la frappe d'une touche, et éventuellement d'une touche simultanée (comme pour produire des majuscules).

- Les caractères initialement affectés aux codes réutilisés ne peuvent plus être affichés sans changer de police de caractères[6].
- Elle complique l'échange de textes car, en l'absence de consensus, de nombreuses polices concurrentes ont été créées : elles sont destinées à la même langue mais n'ont pas affecté les caractères spéciaux aux mêmes codes. La table 1, par exemple, présente 4 polices couramment utilisées pour écrire le bambara. On y voit que le caractère ɔ (o ouvert) occupe respectivement les codes ascii 181, 60 et 249 dans les polices Alphafrica, Arial Bambara et Bambara Arial (dans la table ascii, ces codes étaient initialement affectés aux caractères µ, < et ù). Cette situation a pour conséquence que tout fichier communiqué hors de son lieu de production doit être accompagné des fichiers contenant la ou les polices nécessaires à son affichage.
- Elle interdit tout traitement informatique des textes sans une normalisation préalable car un unique code peut représenter des lettres différentes, or un processus informatique ne voit pas les lettres, il les perçoit à travers le code qu'elles utilisent[7].

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

Table 1 : Encodage de caractères spéciaux du bambara dans 4 polices de caractères

Comme le souligne justement Patrick Andries « Il convient de renoncer définitivement à l'encodage en ISO latin-1 (ISO-8859-1) à l'aide des vieilles polices True Types (SIL ou autre) et de passer au codage en Unicode, qui offre le net avantage de proposer une table de caractères

[6] La police 'Albasa Tamajaq' est dédiée à l'affichage textes en langue tamajak (langue des touaregs). Cette langue utilise 11 caractères spéciaux, doublés de leurs versions majuscules qui occupent donc 22 codes. C'est ainsi que le 'Ǝ' (lettre e majuscule renversée) occupe le code 80, celui du P car la lettre P ne figure pas dans la langue tamajak. Dans un texte sur le football nous avons rencontré le nom de 'Pelé' que seule une lecture attentive du texte a permis de détecter car il était affiché 'Ǝelé'.

[7] Ainsi les polices Arial Bambara et Alphafrica, toutes deux couramment utilisées pour écrire le bambara, affectent le code 249 aux caractères 'Ɛ' (epsilon majuscule) et 'ɔ' (o ouvert) respectivement.

unique proposant l'ensemble des caractères nécessaires à la représentation d'un grand nombre de polices africaines. »

- **Interface humain vers machine**

Il s'agit de la production de textes, leur saisie sous un format électronique.

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, serer et wolof (http://www.termisti.refer.org/ltt/ltt.htm). Ces claviers permettent d'obtenir les caractères spéciaux requis par la frappe d'une seule touche, le code généré est le code Unicode du caractère. Il est facile de matérialiser les caractères sur un clavier physique en posant un cache comportant ces caractères.

Le développement, et la diffusion de ces claviers constitue un progrès significatif dans la production de textes :

- la tâche de saisie s'effectue dans des conditions ergonomiques satisfaisantes,
- le texte saisi est directement encodé selon le standard Unicode.

- **Logiciels**

Les éditeurs de textes couramment disponibles (comme Word ou 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 mais des difficultés imprévues surgissent alors.

Tout d'abord, il est difficile d'utiliser un logiciel si l'on ne maîtrise pas la langue dans laquelle est rédigée son interface, cette situation réduit donc le nombre d'utilisateurs potentiels. De plus, cette situation est inconfortable : elle oblige l'utilisateur à 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 choisies pour la rédaction dans l'autre langue. Ensuite, les fonctionnalités linguistiques complémentaires, comme la correction automatique de l'orthographe, sont inutilisables dès que l'on change de langue. Bien qu'imparfaits, il est évident que ces outils linguistiques sont un soutien appréciable pour produire des textes de qualité. Enfin, les éditeurs de texte produisent, dans le meilleur des cas, des textes codés en ascii, ils n'offrent pas de possibilité de coder ces textes selon le standard Unicode.

Pour l'instant, il n'existe pas d'éditeurs développés pour les langues moins dominantes, comme les langues africaines. Les progrès réalisés dans le codage des caractères, dans l'affichage des textes, et dans la mise au point de claviers virtuels permettent d'envisager le développement de tels logiciels.

1.2. Statut des langues africaines

- **Transcription**

On estime qu'il existe plus de 2000 langues en Afrique. L'écriture de ces langues a généralement emprunté des systèmes de transcription exogènes. Certaines langues ont été écrites avec les caractères arabes avant la colonisation européenne, puis avec les caractères romains. Ces deux systèmes d'écriture ne sont pourtant pas toujours adéquats, par exemple les caractères romains ne permettent pas de représenter correctement les sons des mots arabes du kiswahili [Owino 2002]. De plus, les transcriptions des colonisateurs, réalisées par des amateurs, ont été largement influencées par leur langue d'origine. Par exemple, le wolof est une langue parlée à la fois au Sénégal, colonisé par les français, et en Gambie, colonisée par les anglais. Le son [nja :y] est transcrit *ndiaye* au Sénégal, et *njie* en Gambie : 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].

Les indépendances acquises, les jeunes états africains ont décidé de politiques linguistiques volontaires. Dès 1966 se tint à Bamako sous l'égide de l'UNESCO une réunion qui élabora des alphabets pour les principales langues de l'Afrique de l'Ouest. Un alphabet harmonisé fut ainsi proposé pour le mandingue[8]. Cet alphabet était tout à fait remarquable et certainement le meilleur que l'on puisse imaginer. Par exemple, il notait l'accent aigu sur les mi-fermés **é** et **ó,** et non l'accent grave sur les mi-ouvertes **è** et **ò**, ce qui permettrait une certaine harmonisation avec le mandingue-ouest du Sénégal et de la Gambie. Néanmoins, il ne fut adopté nulle part, car il choquait certaines habitudes et n'était pas harmonisé avec les autres alphabets proposés pour les autres langues (peul et songhaï notamment). Des alphabets différents furent créés dans les divers Etats mandingophones, chacun divergeant d'une manière ou d'une autre de l'alphabet de Bamako.

En juillet 1978, se tint à Niamey dans le cadre du CELTHO (Centre d'études linguistiques et historiques par tradition orale) une nouvelle réunion organisée par l'UNESCO qui créa un « Alphabet africain de référence » , fondé sur les conventions de l'IPA (International Phonetic Association) et de l'IAI (International African Institute). Cet alphabet visait à noter toutes les possibilités phoniques des langues africaines. Par la suite, en novembre de la même année se

[8] Famille de langue dont fait partie le bambara

tint encore à Niamey une « Réunion sur l'harmonisation de l'orthographe du manden » qui visait à harmoniser les différents alphabets mandingues qui s'étaient développés dans les années 1970 de manière divergente. D'autres problèmes furent également étudiés au cours de cette réunion : les tons, la segmentation et l'élision, la ponctuation, etc.

Dans le cadre du projet MAPE, c'est de nouveau l'alphabet de Niamey (novembre 78) qui fut retenu pour l'harmonisation. Les problèmes d'orthographe furent discutés lors de quatre réunions : Abidjan (déc. 80), Bamako (juin 81), Nouakchott (novembre 81), Ouagadougou (juin 82). Des règles de notation des tons, améliorant les propositions antérieures, furent adoptées à la réunion de Bamako de juin 81.

Au cours des années 1980, au Mali, l'ancien alphabet fut peu à peu remplacé par le nouveau système, malgré de nombreuses résistances. Pendant longtemps le nouveau système resta théorique. Ce n'est que vers 1988, lorsque la DNAFLA (Direction Nationale de l'Alphabétisation Fonctionnelle et de la Linguistique Appliquée) acquit des logiciels possédant les nouveaux caractères que commença à se généraliser le nouvel alphabet. Ce nouvel alphabet qui est actuellement essentiellement utilisé au Mali, a permis de résoudre certains problèmes ; mais il en a créé quelques autres. Par ailleurs, il n'a pas été intégralement adopté en Côte d'ivoire et au Burkina Faso.

En Guinée, c'est lors du « Séminaire sur la réforme du système de transcription des langues guinéennes « (27 juillet–2 août 1988) qu'il fut décidé de remplacer l'ancien alphabet par le nouveau. Au Burkina Faso et en Côte-d'Ivoire, l'alphabet utilisé est identique au système en vigueur au Mali, la seule différence est le remplacement de la lettre 'ɲ' par '**ny**'.

A côté de tous les travaux qui sont faits actuellement pour développer l'écriture du mandenkan (et des autres langues africaines) en caractères latins, il existe d'autres projets dans des systèmes différents, essentiellement l'alphabet arabe et le Nko.

Il y a eu des tentatives pour transcrire les langues africaines avec les caractères arabes, patronnées par l'UNESCO ; mais actuellement ce projet est essentiellement soutenu par l'ISESCO (Islamic Education Science and Culture Organisation). L'idée de l'ISESCO est de créer une transcription arabe moderne pour les langues africaines en se dotant sur la tradition graphique adjami et l'expérience des langues de l'Asie musulmane. Cette transcription se répand à travers le réseau des écoles coraniques.

Le Nko connaît aujourd'hui un succès considérable à tous les niveaux. Il se diffuse notamment par le milieu des écoles coraniques et des commerçants. Ce système a été inventé par en 1947 par Souleymane Kanté, érudit dont les multiples connaissances couvrait aussi bien l'histoire traditionnelle, la pharmacopée que l'étymologie du vocabulaire mandingue. A l'origine l'alphabet Nko avait été conçu pour le mandingue (où « je dis » se dit N ko dans tous les dialectes), mais il

est maintenant utilisé pour d'autres langues africaines. « Malgré leurs qualités indéniables, l'on peut craindre que le système adjami de l'ISESCo et le système Nko créent des divisions supplémentaires dont l'Afrique n'a pas besoin. » (Gérard Galtier)

Cette histoire de la transcription témoigne de la conscience des états de l'Afrique de l'Ouest de l'importance de bien écrire ses propres langues. La faiblesse économique de ces pays, conjuguée au manque de cadres qualifiés, a malheureusement entravé la production des savoirs linguistiques nécessaires (dictionnaires, grammaires, etc.). Ainsi, la plupart des langues ne bénéficient d'aucun dictionnaire monolingue[9], ce qui constitue une situation paradoxale puisque qu'elles sont souvent dotées de plusieurs dictionnaires bilingues. Ces ressources ne sont quasiment pas distribuées, car les ouvrages restent trop onéreux. Au nom du principe de réalité, ce problème de coût a poussé les autorités locales à favoriser la production de petits manuels, généralement destinés aux formateurs.

- Enseignement

En Afrique de l'Ouest, le défaut d'alphabétisation de la population dans les langues africaines a longtemps prévalu. Historiquement, les colonisateurs ont imposé l'enseignement dans leur langue, langue devenue officielle après les indépendances afin de ne pas favoriser une langue par rapport aux autres. La post-colonisation a vu les élites africaines prolonger cette situation, car c'est généralement la langue officielle[10] du pays, et les décideurs, alphabétisés dans la langue du colonisateur, sont quasiment analphabètes dans leur langue maternelle qu'ils n'ont pas appris à lire, et encore moins à écrire.

Cependant, au Mali, la réforme de 1962 préconisait déjà l'utilisation des langues nationales dans l'enseignement dès que les moyens le permettraient. Des mesures ont alors été prises pour entreprendre les recherches fondamentales sur les langues nationales. Suite à l'adhésion du Mali au concept d'alphabétisation fonctionnelle dans les langues nationales défini par la Conférence de Téhéran en 1965, quatre langues sont utilisées comme médiums d'enseignement (bambara, peul, songhaï et tamasheq cf. [Calvet 1984 :112]).

Suite à une forte recommandation faite par le deuxième séminaire national sur l'éducation en 1978, les langues nationales instrumentalisées sont introduites comme médiums et matières dans le système formel au niveau de quelques écoles cibles du premier cycle de l'enseignement fondamental de 1979 à 1993. La méthodologie convergente d'apprentissage des langues nationales et du français expérimentée à Ségou en 1987 a connu un succès certain, succès qui a conduit à sa généralisation progressive à toutes les écoles à partir de 1994. Elle constitue

[9] Nous pouvons citer un dictionnaire monolingue zarma : Isufi Alzuma Umaru, "Kaamuusu Kayna", éd. Alpha, 1996.

[10] La langue de l'Etat

aujourd'hui le socle du curriculum de l'enseignement fondamental dans le cadre de la nouvelle politique éducative du Mali à travers le PRODECC.

En 1993, une autre alternative éducative s'est fait jour : les Centres d'Education pour le Développement (CED) ciblent les enfants de 9 à 15 ans, non scolarisés ou déscolarisés du système formel. Ils utilisent les langues nationales comme médiums d'apprentissage concomitamment avec le français.

2. Objectifs

2.1. Favoriser la production d'écrits

Conscients de l'enjeu vital que représente l'alphabétisation de la population, les Etats d'Afrique de l'Ouest ont mis en œuvre une politique de planification linguistique importante, choisissant leurs langues nationales, normalisant les alphabets et les règles de transcription, soutenant des centres de production de livres d'alphabétisation, de guides médicaux, de manuels divers (gestion, droits des enfants, récits). Pourtant, les participants à la récente conférence nationale sur la promotion des langues au Mali[11], ont déploré le manque d'écrits en langue nationale : les personnes alphabétisées dans leur langue perdent peu à peu leurs connaissances car elles n'ont quasiment jamais l'occasion de lire, et ce qu'il y a à lire ne correspond pas toujours à leurs centres d'intérêts. En particulier il y a peu de livres pour enfants, quasiment aucune bande dessinée. Il apparaît donc vital d'encourager et de soutenir la production d'écrits en langues nationales.

Nous faisons l'hypothèse que les personnes rédigeant des textes en langue nationale (manuels scolaires, journaux, manuels techniques, sites Internet) pourraient maintenir une bonne qualité de langue et produire des textes en plus grande quantité si elles avaient accès à :

- des ressources linguistiques (lexiques, dictionnaires, grammaires) leur permettant de maintenir leur maîtrise de l'écrit dans leur propre langue (dans laquelle elles excellent oralement).
- des textes, nombreux, variés et de bonne qualité linguistique.

2.2. Production et distribution de ressources linguistiques

Les obstacles à la production et à la distribution de ressources linguistiques sont nombreux. Il est très difficile de produire des ressources linguistiques importantes comme les dictionnaires, or les personnes qualifiées sont rares en Afrique de l'Ouest et généralement déjà mobilisées sur des tâches également importantes comme la production de manuels d'éducation ou de santé. L'impression et la distribution représentent également un frein important car les moyens sont

[11] Bamako, 15-17 janvier 2004

limités (et de tels les ouvrages imprimés restent de toute façon trop coûteux pour être réellement accessibles).

L'utilisation des ordinateurs pour la saisie des textes représente une possibilité de pallier certaines difficultés. La plupart des grandes villes disposent maintenant de cybercafés où le coût de connexion à Internet est assez modeste, il est facile d'y télécharger des fichiers et des logiciels. La diffusion et la mise à jour de ressources linguistiques grâce à Internet est une possibilité tout à fait réaliste. Dans ce nouveau contexte, la production exhaustive de ressources linguistiques classiques peut être remplacée par la mise au point progressive de ressources électroniques. Ces ressources peuvent être de différentes natures : utilisables directement par des humains (dictionnaires en ligne) ou par les logiciels qu'ils utilisent pour écrire. Ainsi, un éditeur de textes adapté à la langue de l'utilisateur peut fournir un soutien appréciable à l'écriture de textes. Un tel éditeur doit en outre produire des fichiers codés selon le standard Unicode, et offrir un environnement ergonomique de haute qualité.

Cette question de l'ergonomie est fondamentale car la généralisation de l'outil informatique entraîne une modification des pratiques. D'abord utilisés majoritairement par des secrétaires pour effectuer la saisie de textes écrits au préalable, les ordinateurs servent de plus en plus à rédiger directement les textes. Pouvoir s'abstraire des contraintes techniques est essentiel à la concentration qu'exige la formulation ex nihilo de textes écrits. Dans ce cas, l'ergonomie réside surtout dans l'effort intellectuel minimal à fournir pour former les lettres composant le texte, le modifier, le mettre en forme, etc. Le clavier doit donc être adapté à la production de tous les caractères requis. Il est également important que la langue de fonctionnement du logiciel soit celle de l'utilisateur afin d'éviter la gymnastique intellectuelle nécessaire pour écrire dans un environnement bilingue.

2.3. Conséquences

L'apparition de logiciels adaptés à la production de textes électroniques en langues africaines, en favorisant la production d'écrits de bonne qualité autant linguistique que technique (Unicode), va permettre d'utiliser d'autres outils informatiques pour mieux étudier ces langues. La capitalisation de textes électroniques constitue, en effet, une mine d'informations. Certaines applications sont directement envisageables. Ainsi, la mesure des fréquences des mots dans un corpus représente un nouveau critère pour décider quels mots doivent figurer dans un lexique de base, ou dans un dictionnaire. Un concordancier représente une aide précieuse qui permet de distinguer les différentes significations d'un mot, ou de comparer les contextes d'utilisation de deux mots. Des textes de variantes dialectales peuvent être comparés statistiquement. D'autres développements à moyen terme peuvent également être envisagés, comme l'extraction automatique de termes, ou l'aide à la traduction.

3. Réalisation

3.1. Correcteur orthographique

- **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 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 indetectées augmente avec l'accroissement de la taille du lexique. 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 (généralement via des calculs statistiques) peut aider à éviter cet écueil majeur.

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 sélectionnés 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 adapter 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 dans les 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 existants auparavant.

- Inadéquation des correcteurs orthographiques existants pour les langues africaines

Il existe déjà des correcteurs orthographiques destinés à 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 se heurter à une limitation inévitable de 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 habituel : dans la grande majorité des cas, l'utilisateur ne dispose pas de ressources linguistiques imprimées (dictionnaire) lui permettant de lever des questions sémantiques (pour vérifier le sens d'un mot par exemple) ou syntaxiques. Par ailleurs, ces ressources sont quasi-inexistantes.

Nous pensons qu'un correcteur orthographique qui, par définition, traite des textes, se situe à une place stratégique pour dispenser des ressources linguistiques à l'utilisateur, et pour aider à la constitution de ressources linguistiques.

- 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. Nous avons défini des fonctionnalités supplémentaires pour, d'une part, communiquer d'avantage d'informations linguistiques à l'utilisateur, 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). La communication d'informations supplémentaires sur les mots (comme sa ou ses définitions, sa catégorie lexicale, 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.

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. Les ressources sont rassemblées dans le lexique électronique de la langue. Chaque item y est, dans la mesure du possible, accompagné de sa catégorie grammaticale, de son mode de flexion et des dérivations possibles afin de pouvoir étendre le lexique à toutes les formes calculables (nous détaillons en annexe les différentes dérivations lexicales en bambara). Les informations contextuelles sont également intégrées au lexique sous forme de probabilités. Pour adapter ce correcteur orthographique à une nouvelle langue, il suffit donc de changer les ressources lexicales.

Le correcteur orthographique lui-même est en développement à l'Université de Nantes, il sera compatible avec les éditeurs de textes couramment utilisés (comme Word) ainsi que les éditeurs de messages électroniques. Le développement de la version pilote devrait être achevé en 2004. Les tests réalisés début 2005 déboucheront sur la réalisation de la version finale. Celle-ci devrait être disponible, et gratuitement téléchargeable, au cours de l'année 2005.

- **Soutien à la production de ressources linguistiques**

Nous avons déjà souligné le manque de ressources linguistiques sur les langues africaines. Il est capital favoriser leur développement, car ces ressources constituent la base des logiciels de correction orthographique. Les travaux à mener étant très importants, et les personnes qualifiées plutôt rares et déjà surchargées de travail, il faut donc fournir des outils d'aide à la production de ces ressources.

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 d'un correcteur orthographique ;
- enrichissement des items du lexique (catégorie lexicale, définitions, exemples d'usage, etc.) ;
- calcul automatique d'informations statistiques (trigrammes sur les symboles, fréquence d'apparition des mots, etc.).

Les personnes chargées de la maintenance des ressources électroniques verraient leur tâche facilitée par l'utilisation d'un 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. Ces informations peuvent être directement inscrites dans le lexique électronique de la langue.

3.2. Constitution de ressources linguistiques sur le bambara

Au Mali, de nombreux textes sont produits dans les 11 langues nationales. Nous avons choisi de commencer par travailler sur le bambara, langue véhiculaire dominante dans le sens où elle est souvent utilisée comme langue de substitution par deux maliens de langues maternelles différentes [Calvet 1981]. La constitution de ressources linguistiques électroniques est en cours, elle se déroule en plusieurs étapes :

- recueil de textes variés (articles de journaux, manuels de santé, de psychologie, de gestion, contes, dictionnaire, etc.),
- conversion de ces textes, de format électroniques variés, au format Unicode,
- représentation normalisée des textes à l'aide d'XML.

En janvier 2004, nous avons collecté à Bamako un corpus de 89 684 mots en bambara. Ces textes sont issus de plusieurs auteurs (journalistes, écrivains) et appartiennent à différents genres (contes, manuels techniques, manuels de santé, récits, etc.). Ce corpus n'est évidemment pas définitif et sera complété au gré des nouvelles trouvailles.

La première étape pour l'exploitation d'un tel corpus est sa normalisation. Les textes sont tous écrits en bambara correct, respectant les décrets en vigueur, mais ils ont été produits avec des

polices « redessinées » pour afficher les caractères spéciaux, ce qui rend impossible leur exploitation électronique tels quels. Nous avons donc modifié le codage de ces caractères spéciaux afin de respecter le standard Unicode, et avons balisé les textes avec des balises XML[12] conformément la norme XCES[13].

Nous avons aussi recueilli les fichiers électroniques constituant un dictionnaire bilingue bambara-français d'environ 8000 entrées [Bailleul 1996]. Ce dictionnaire constitue une importante ressources car les catégories lexicales des mots (nom, verbes), sont indiquées, ce qui permet d'envisager le calcul automatique des flexions et dérivations des mots en sein du correcteur orthographique. La transcription des mots diffère, cependant, des règles officielles car le bambara est une langue tonale et agglutinative ; le dictionnaire étant a priori destiné à des étrangers, l'auteur a indiqué les tons sous forme de signes diacritiques et il a rendu visible l'agglutination de mots en la marquant par un point et en donnant la traduction de chacun des mots, puis la traduction du composé (exemple : fàri.gan [corps.chaud] = fièvre). Des programmes ont donc été élaborés pour extraire les entrées du dictionnaire et leur catégorie grammaticale, modifier leur transcription (en supprimant signes diacritiques et point) afin de respecter les décrets. Le lexique a finalement été balisé conformément à la norme XCES.

Le corpus de textes ainsi que le dictionnaire vont permettre d'ébaucher le lexique électronique du bambara. Tous les mots des textes seront intégrés à ce lexique, ainsi que les mots du dictionnaire auxquels il faudra ajouter leurs flexions. Quelques calculs statistiques pourront être mis en place, bien que la taille du corpus soit insuffisante pour qu'ils soient vraiment significatifs. Nous pourrons augmenter la taille du corpus en lui ajoutant des textes afin d'améliorer la fiabilité de ces calculs.

4. Conclusion

Les obstacles techniques à l'écriture des langues africaines disparaissent grâce à l'émergence du standard Unicode pour le codage des caractères spéciaux. Les outils informatiques ne sont pas toujours adaptés à ces langues, mais d'ici quelques années des progrès considérables devraient avoir lieu, permettant ainsi aux africains d'utiliser leurs propres langues dans les environnements de communication les plus modernes.

L'apparition de correcteurs orthographiques adaptés aux langues africaines devrait favoriser la production d'écrits et augmenter la présence des langues africaines sur Internet. Le développement effectif d'un correcteur pour le bambara ainsi que la constitution de ressources linguistiques électroniques pour cette langue constituent un projet pilote destiné à valider

[12] eXtended Markup Language http://www.w3.org/XML/
[13] Corpus Encoding Standard for XML http://www.xml-ces.org/

l'approche que nous avons décrite. D'autres travaux suivant la même stratégie sont initiés dans d'autres langues de la région (wolof au Sénégal, hausa et zarma au Niger).

5. Bibliographie

[Andries 2002] Andries, P., Introduction à Unicode et à l'ISO 10646, Document numérique, vol.6, n°3-4, pp.51-88, 2002.

[Bailleul 1996] Bailleul, C. "Dictionnaire bambara-français", éd. Donniya, Bamako, Mali, 1996.

[Calvet 1981] Calvet, L.-J., "Les langues véhiculaires", PUF. Que sais-je ?, 1981.

[Calvet 1984] Calvet, L.-J., "La tradition orale", PUF. Que sais-je ?, 1984.

[Chanard 2001] Chanard, C., Popescu-Belis A., "Encodage informatique multilingue : application au contexte du Niger". Les cahiers du RIFAL n°22 , pp.33-45, 2001.

[Kukich 1992] Kukich, Karen. "Techniques for automatically correcting words in text". ACM *Computing Survey*s, 24(4), pp.377–439. 1992.

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

[Owino 2002] Owino, F. R., "The expansion of dholuo vowel system", Writing African – The Harmonisation of Orthographic Conventions in African Languages, ed. Kwesi Kwaa Prah, pp. 151-160, 2002.

[Van der Veken 2003] Van der Veken, A., de Schryver, G.-M., "Les langues africaines sur la Toile : études des cas haoussa, somali, lingala et isixhosa". Les cahiers du RIFAL n°23, pp.33-45, novembre 2003.

Annexe : Les affixes en bambara

- **Les préfixes**

Tous se rattachent à une base verbale.

préfixe	marque	exemples
la–	causalité	la– + bɔ (sortir) = labɔ (faire sortir) la– + kuma (parler) = lakuma (faire parler)
ma–	contact	ma– + bɔ = mabɔ (écarter, éloigner) ma– + jigin (descendre) = majigin (l'amener vers soi)
sɔ–	contact	sɔ– + bɔ = sɔbɔ (écarter, éloigner sɔ– + don (entrer) = sɔdon (s'approcher)

- **Les suffixes de dérivation**

Suffixes de dérivation se rattachant à une base nominale

suffixe	marque	exemples
–la[14]	lieu	Kulubali (nom de famille) + –la = kulubalila (chez les Coulibaly) cɛ (mari) + –la = domicile conjugal kɔnɔ (ventre) + –na = kɔnɔna (intérieur)
–ka	habitant de	bamakɔ + –ka = bamakɔka (qui vient de Bamako, Bamakois)
–ya[15]	état	mɔgɔ (l'humain) + –ya = mɔgɔya (le fait d'être humain) muso (femme) + –ya = musoya (le fait d'être femme, la féminité)
–tɔ	défaut, maladie	fiyen (la cécité) + –tɔ = fiyentɔ (qui est atteint de cécité, aveugle)
–ma	qui contient	kɔgɔ (sel) + –ma = kɔgɔma (qui a du sel, salé)
–ntan	privation	kɔgɔ (sel) + –ntan = kɔgɔntan (qui n'a pas de sel, non salé)
–lama[16]	aspect	mɔgɔ + –lama = mɔgɔlama (qui a l'aspect d'une personne)
–ba	augmentatif	wari (argent) + –ba = wariba (une grande somme)
–nin	diminutif	wari + –nin = warinin (une petite somme)
–w[17]	pluriel	so (maison) + –w = sow (maisons)

Suffixes de dérivation se rattachant à une base verbale et formant avec les bases auxquelles ils se rattachent des dérivés nominaux ;

suffixe	rôle	exemples

[14] devient –na après une consonne contenant une nasale

[15] le suffixe –ya peut aussi se rattacher à une base adjectivale :
surun (court) + –ya = surunya (le fait d'être petit, la petitesse)
dun (profond) + –ya = dunya (le fait d'être, profondeur)

[16] devient –nama après une consonne contenant une nasale

[17] le pluriel peut être marqué par le suffixe –lu (–nu après une consonne contenant une nasale) dans deux cas
o (il, elle) + –lu = olu (ils, elles)
nin (démonstratif) + –nu = ceux-ci
Ces suffixes forment avec le radical des dérivés nominaux.

–la[18]	agent habituel d'une action	dununfɔ (battre le tam-tam) + –la = dunufɔla (batteur de tam-tam) misigɛn (chasser les bœufs) + –na = misigɛnna (berger, pasteur)
–baga (ou –baa)	agent ponctuel d'une action	dɛmɛ (aider) + –baga = dɛmɛbaga (qui aide)
–li[19]	action	taa (partir) + –li = taali (le fait de partir, départ) segin (retourner) + –ni = retour
–lan[20]	instrument	sɛbɛnnikɛ (écrire) + –lan = sɛbɛnnikɛlan (instrument avec lequel on écrit) misigɛn + –nan = misigɛnnan (instrument avec lequel on chasse les bœufs)
–nci	superlatif ou excès (souvent péjoratif)	janfa (trahir) + –nci = janfanci (traître) hinɛ (pitié) + –nci = hinɛnci (trop miséricordieux)
–bali	négation , privation	taa (partir) + –bali = taabali (le fait de ne pas partir)
–ta	destination	dun (manger) + –ta =dunta (destiné à être manger, qui est à manger)

Suffixe se rattachant à une base adjectivale et formant avec le radical des dérivés adjectivaux.

suffixe	marque	exemples
–man	qualité	kɛnɛ (en santé) + –man = kɛnɛman (bien portant)

Suffixes se rattachant à une base numérale et formant avec le radical des dérivés numéraux.

suffixe	marque	Exemples
–nan	ordre, rang	fila (deux) + –nan = filanan (deuxième)
–la[21]	quantité	saba (trois) + –la = sabala (qui coûte quinze francs, 3 fois 5 francs)

- **Les suffixes de conjugaison**

Suffixes rattachés à une base verbale pour indiquer très généralement un passé : –ra[22], –len[23] et tɔ.

verbe	Exemple
ka taa (partir)	Cɛw ni musow taara. (les hommes et les femmes sont partis.)

[18] devient –na après une consonne contenant une nasale

[19] devient –ni après une consonne contenant une nasale

[20] devient –nan après une consonne contenant une nasale

[21] devient –na après une consonne contenant une nasale

[22] devient –la après une syllabe consonne contenant 'l' ou 'r' et –na après une syllabe contenant une nasale

[23] devient –nen après une consonne contenant une nasale

ka wasa (être satisfait)	Bɛɛ wasara. (tout le monde est satisfait.)
ka boli (courir)	U bolila kunun. (Ils ont couru hier.)
ka wuli (se lever)	Denmisɛnw wulila. (Les enfants se sont levés.)
ka segin (revenir, retourner)	Taamadenw seginna. (Les voyageurs sont revenus.)
ka kuma (parler)	An kumana kosɛbɛ. (Nous avons beaucoup parlé.)
ka taa (partir)	A taalen, jama wulila. (la foule s'est levée quand elle est partie
ka taa (partir)	Taamadenw seginnen, a taara. (Il est parti quand les voyageurs ont été de retour.)
ka bin (tomber)	A taatɔ binna. (Il est tombé en partant.)