Des correcteurs orthographiques pour collecter et diffuser les connaissances linguistiques en Afrique subsaharienne

Chantal Enguehard

Des correcteurs orthographiques

pour collecter et diffuser les connaissances linguistiques

en Afrique subsaharienne

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

1. Introduction

Les pays du Sahel figurent parmi les plus pauvres de la terre, ils connaissent aussi des taux d'analphabétisation particulièrement élevés (83% au Niger, 60% au Mali, 57% au Sénégal). Or, l'alphabétisation de la population est un facteur essentiel pour développer un pays, il apparaît donc essentiel de mettre en place des stratégies pour améliorer l'alphabétisation des populations.
Nous passons rapidement en revue l'histoire récente de la transcription des langues africaines et les problèmes rencontrés dans la production et la diffusion de connaissances linguistiques et de textes. Ensuite nous proposons le développement d'outils informatiques spécifiques pour capitaliser des connaissances et les diffuser.

2. Situation

2.1. Bref rappel historique

la période coloniale

La colonisation des pays africains a mis en péril des systèmes de transcription émergents qui existaient déjà, et notamment 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, les syllabaires mandé qui apparaissent entre 1833 et 1930, l'alphabet Nko ou encore l'écriture arabe [Dalby 1986]. 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[1]. Il faut souligner que ces travaux d'amateurs ont fini par imposer l'utilisation de l'alphabet latin.

Les colonisateurs ont eu des comportements différentes en ce qui concerne les langues locales dans l'enseignement. Dans l'ex-Congo belge (aujourd'hui RDC[2]), certaines langues africaines étaient enseignées alors qu'au même moment elles n'avaient aucun droit de cité dans les colonies françaises.

La scolarisation dans les langues européennes n'a pas été un succès. Ainsi, dès 1817, Jean Dard [Gaucher 1968], un instituteur français en poste à l'école mutuelle de Saint-Louis du Sénégal dont il était le directeur, constatant « l'échec d'un enseignement unilingue français amenant les élèves à lire et à écrire le français sans le comprendre » [Dumont 1983] tente d'utiliser le wolof comme langue d'enseignement. Le 25 mars 1829, la commission scolaire envoyée par le gouverneur Dubelin pour contrôler l'école condamna et mit fin à cette innovation pédagogique.

Il faut bien constater que cette politique assimilationniste de la France d'alors 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 chronique. Comment apprendre à lire et à écrire dans une langue étrangère que l'on ne comprend pas, que l'on ne côtoie pas ?

La postcolonisation

Très tôt, les intellectuels africains ont compris que l'alphabétisation de la population, facteur essentiel pour

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

2 République Démocratique du Congo.

développer un pays, doit nécessairement s'appuyer sur l'enseignement des langues maternelles[3]. Après les décolonisations, des efforts considérables ont donc été menés pour transcrire les langues africaines, mais le retard en la matière, associé à la confusion introduite par les linguistes amateurs ont considérablement compliqué la tâche. Il faut bien mesurer la différence de situation linguistique entre de grandes langues internationales comme le français ou l'anglais, sur lesquels des milliers de linguistes ont fait des études très poussées, qui bénéficient d'académies nationales des langues depuis plusieurs siècles, et celle des langues africaines dont le passage historique naturel de l'oral vers l'écrit a été brutalement interrompu puis largement corrompu par des éléments exogène. Les travaux à mener sur les langues africaines représentent une tâche immense qu'il est crucial de mener à bien afin d'en faire bénéficier la population en termes de développement et d'éducation.

Depuis 50 ans, de nombreux travaux de grande importance ont été menés, mais la linguistique est une science difficile car elle concerne l'ensemble des locuteurs, il faut du temps (et des moyens) pour que ses travaux soient diffusés, pénètrent les populations et soient finalement adoptés.

Les travaux pour définir les alphabets (généralement sous l'égide de l'UNESCO) reflètent bien ces difficultés : l'alphabet défini pour les langues mandingue[4] à Bamako en 1966, bien que remarquable puisqu'il permettait une certaine harmonisation avec le mandingue-ouest du Sénégal et de la Gambie, choquait certaines habitudes et n'était pas harmonisé avec les autres alphabets proposés pour les autres langues (peul et songhaï notamment). Pendant la décennie suivante, des alphabets différents ont été créés dans les divers états mandingophones, chacun divergeant d'une manière ou d'une autre de l'alphabet de Bamako. En juillet 1978, une nouvelle réunion a abouti à la création d'un « Alphabet africain de référence » fondé sur les conventions de l'IPA (International Phonetic Association) et de l'IAI (International African Institute). Cet alphabet permet de noter toutes les possibilités phoniques des langues africaines. Enfin, quelques mois plus tard, lors des travaux sur l'harmonisation de l'orthographe du manden (différents alphabets mandingues qui s'étaient développés dans les années 1970 de manière divergente) de nouveaux problèmes sont également abordés : les tons, la segmentation et l'élision, la ponctuation, etc. L'alphabet adopté est encore amélioré en juin 1981 par des règles de notation des tons.

Il aura donc fallu une quinzaine d'années pour définir un alphabet adéquat. Comme cet alphabet comporte des caractères ne figurant pas dans l'alphabet latin (par exemple ɛ, ɔ, ɲ, ŋ), il a rencontré de grandes résistances car il posait des problèmes techniques pour la saisie de textes, leur édition, etc., et il semblait plus simple de continuer à n'utiliser que l'alphabet latin occidental. Ainsi, en Guinée, il n'est adopté que depuis 1988. Au Burkina Faso et en Côte-d'Ivoire, il n'est encore que partiellement utilisé puisque la lettre 'ɲ' est remplacée par 'ny'.

2.2. Evolution technique

Comme les caractères spéciaux des alphabets africains ne figurent pas dans la représentation ascii, de nombreuses polices de caractères ont été créées lors des années 80 : les glyphe de quelques caractères inutiles (&, @, <, etc.) y sont redessinés pour figurer les caractères spéciaux. Aucune harmonisation n'ayant prévalu, les caractères spéciaux ne remplacent pas toujours les mêmes caractères initiaux, rendant quasiment impossible l'échange de documents, et leur traitement électronique. Ainsi des textes en 7 langues recueillis en 2004 au Niger et au Mali n'utilisent pas moins de 25 polices de caractères différentes [Enguehard 2004] !

L'apparition du standard Unicode en 1992 représente un progrès considérable puisqu'il permet de représenter plus d'un million de caractères différents, c'est-à-dire tous les caractères de toutes les langues. Ce nouveau codage des caractères 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].

Cependant, quelques problèmes subsistent (voir les illustrations Table 1) :

Cas 1 : Certains caractères sont encore absents, c'est le cas du 'r' avec une barre horizontale, (langue kanuri[5] - arrêté 213-99 de la République du Niger). Cette lettre peut être composée par la superposition d'un tiret '-' à la lettre 'r' mais, comme il n'est pas possible de spécifier la position d'un signe par rapport à l'autre, le rendu final n'est pas satisfaisant. De plus, les utilisateurs sont tentés d'utiliser des tirets de codages différents selon le rendu obtenu (tiret long, tiret court), ce qui nuit à une représentation unique de ce caractère.

Cas 2 : D'une manière générale, de nombreux caractères avec signes diacritiques ont plusieurs codages possibles.

3 Lors de sa conférence à la Chambre de commerce de Dakar, en 1937, le futur Président de la République, Léopold Sédar Senghor, préconisa l'enseignement des langues maternelles.

4 Famille de langue dont fait partie le bambara

5 Une des langues nationales de la République du Niger

Ainsi ‘e’ avec accent aigu est représenté par le code U+00E9, mais également par la combinaison de la lettre ‘e’ et du code de l’accent aigu : U+0065 U+0301. Ces multiples codages compliquent le traitement analytique des textes. De plus il arrive que les logiciels affichent différemment ces caractères accentués, selon le codage utilisé. En particulier, un caractère présentant déjà un signe supplémentaire (comme le point au-dessus du ‘i’), superposé à un signe diacritique, est affiché en superposant le point et le signe diacritique, ce qui est incorrect : le point au-dessus du ‘i’ devrait être remplacé par le signe diacritique. Ainsi, le caractère ‘j avec caron’ (langue tamasheq[6] - arrêté 214-99 de la République du Niger) est codé par l’unique code U+01F0 sous sa forme minuscule. La forme majuscule, curieusement, est absente de l’alphabet Unicode. Sa création par superposition du ‘J majuscule’ avec l’accent caron est décevante : selon les logiciels, l’accent caron apparaît comme un accent plat, ou bien il n’est pas situé correctement au-dessus du ‘J’.

Cas 3 : Certaines langues utilisent des digraphes, signes composés de deux lettres mais dont la forme capitale ne fait apparaître que la première lettre en majuscule. Le hausa, par exemple, utilise 9 digraphes (arrêté 212-99 de la République du Niger). Un texte peut mélanger des mots de langues différentes, ou des noms propres, dans lesquels peuvent apparaître les mêmes combinaisons de lettres que ces digraphes alors qu’il s‘agit de lettres séparées. En Français, par exemple, le mot "gyrophare" comprend la suite de lettres "gy" mais il ne s’agit pas du digraphe ‘gy’.

Ces problème ne peuvent être gérés par les polices de caractères. La meilleure solution est la définition de codes spécifiques pour les caractères absents, tous les caractères avec signes diacritiques, et tous les digraphes. Cette solution semble envisageable puisque, dans le standard Unicode, de nombreux codes ne sont pas affectés.

	Langue	Lettre minuscule		Lettre majuscule	
		affichage	codes	affichage	codes
‘r’ avec barre	Kanuri	ɾ	U+0072 (r) U+0335 (tiret court)	R	U+0052 (R) U+0335 (tiret court)
		ɾ	U+0072 (r) U+0336 (tiret long)	R	U+0052 (R) U+0336 (tiret long)
e avec accent aigu	Français	é	U+00E9	É	U+00C9
		é	U+0065 + U+0301	É	U+0065 + U+0301
‘j’ avec caron	Tamasheq	ǰ	U+01F0		
		ǰ	U+006A U+030C	J̌	U+004A U+030C
digraphes	Hausa[7]	fy	U+0066 (f) U+0079 (y)	Fy	U+0046 (F) U+0079 (y)
		gw	U+0067 (g) U+0077 (w)	Gw	U+0047 (G) U+0077 (w)
		gy	U+0067 (g) U+0079 (y)	Gy	U+0047 (G) U+0079 (y)
		kw	U+006B (k) U+0077 (w)	Kw	U+004B (K) U+0077 (w)
		ky	U+006B (k) U+0079 (y)	Ky	U+004B (K) U+0079 (y)
		ƙw	U+00199 (ƙ) U+0077 (w)	Kw	U+00198 (K) U+0077 (w)
		ƙy	U+00199 (ƙ) U+0079 (y)	Ky	U+00198 (K) U+0079 (y)
		sh	U+0073 (s) U+0068 (h)	Sh	U+0053 (S) U+0068 (h)
		ts	U+0074 (t) U+0073 (s)	Ts	U+0054 (T) U+0073 (s)

Table 1 : quelques caractères spéciaux

6 Une des langues nationales de la République du Niger

7 Une des langues nationales de la République du Niger

2.3. Situation linguistique

Peu de linguistes

Le travail de description scientifique des langues africaines est loin d'être achevé et de nombreux travaux essentiels restent encore à mener : les langues présentent de nombreuses variantes régionales, les transcriptions concurrentes d'un même mot sont légion, les systèmes orthographiques officiels ne sont généralement 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[8]. Comme nous l'avons vu, ces études doivent impérativement être menées par des linguistes de métier afin de mettre au point des systèmes de transcription cohérents, susceptibles de fixer les langues locales et d'aider à leur introduction dans le système éducatif.

Mais dans des pays démunis où une très faible partie de la population est éduquée, les linguistes sont extrêmement rares alors que le travail de description linguistique qui reste à mener est gigantesque. Les pays concernés n'ont pas les moyens économiques pour former des linguistes, ou donner les moyens aux linguistes formés de travailler dans de bonnes conditions : la plupart des travaux que j'ai décrits ont bénéficié de l'appui d'organismes internationaux comme l'UNESCO, nationaux (coopération), ou privés comme la Société Internationale de Linguistique.

Il faut ajouter que ces pays font face à de tels problèmes de sous-développement (dans le domaine de l'accès à l'eau potable, à la nourriture, à l'éducation, à la santé), qu'ils n'ont pas eu les moyens de rendre obligatoires les décisions prises en matière linguistique par des mesures législatives coercitives : les systèmes de transcription officiels ne sont pas toujours respectés et les décrets sont généralement laissés pour compte.

Pas de ressources linguistiques

La faiblesse économique endémique, conjuguée au manque de cadres qualifiés, a entravé la production des savoirs linguistiques nécessaires au bon usage d'une langue (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.

Cette situation est aggravée par les occidentaux qui, sans machiavélisme mais avec une inconscience remarquable, contribuent encore de nos jours à brouiller les pistes. Ainsi, les ouvrages de botanique, largement diffusés à travers le monde, proposent en fin d'ouvrages les noms des plantes dans plusieurs langues africaines. Ces noms, transcrits par des botanistes dénués de compétences linguistiques, ne respectent aucunement les règles de transcription de ces langues (la plupart du temps, ils ne sont même pas écrits avec l'alphabet adéquat) [Enguehard 2003]. Plus grave, sur la toile, de nombreux sites amateurs donnent des rudiments dans différentes langues sans que leurs auteurs aient conscience qu'ils transcrivent les mots sans respecter les règles de bonne transcription…. dont ils ignorent l'existence.

Ces connaissances frelatées sont largement diffusées à travers le monde (les ouvrages de botanique sont présents dans les bibliothèques, les sites internet sont ouverts à tous) alors que les ressources de référence, produites par les linguistes africains, ne le sont pas car les ouvrages restent trop onéreux pour leurs concitoyens.

Conscients de l'enjeu vital que représente l'alphabétisation de la population, les Etats d'Afrique de l'Ouest ont pourtant 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) et incluant depuis quelques années l'enseignement des langues nationales. Pourtant, les participants à la récente conférence nationale sur la promotion des langues au Mali[10], 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.

---

8 Ici encore il faut gérer le lourd passé colonial. Ainsi, en wolof, l'énoncé [nja :y] est transcrit *ndiaye* au Sénégal (colonisé par les français), et *njie* en Gambie (colonisée par les anglais). Il faut noter qu'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. Un projet de l'ACCT (Agence de Coopération Culturelle et Technique), maintenant AIF (Agence Intergouvernementale de la Francophonie) 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.

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

10 Bamako, 15-17 janvier 2004

Mais les savoirs linguistiques inculqués à l'école élémentaire ne peuvent se pérenniser. Qui imaginerait écrire correctement sa propre langue en ayant quitté l'école à 14 ans, en ne disposant jamais d'aucun dictionnaire ni grammaire, en croisant rarement des textes écrits ? Dans de telles conditions, écrire une langue africaine constitue un exploit qui est l'apanage de rares élites.

La production de ressources linguistiques de référence et leur diffusion massive constituent donc un enjeu majeur pour permettre aux africains d'acquérir une culture de l'écrit, et échapper ainsi à l'analphabétisme chronique.

3. L'informatique, un atout pour les langues africaines

3.1. Hypothèses

Il est impératif de développer des stratégies à faible coût (humain et financier) permettant de constituer et de diffuser des ressources linguistiques.

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 répondant à leurs besoins.

Or, tout texte, à un moment ou à un autre de son élaboration, est saisi et stocké sur support électronique. Nous faisons l'hypothèse que cette utilisation systématique des ordinateurs représente une opportunité à saisir car ils représentent un lieu où diffuser et recueillir des données linguistiques et offrent des outils modernes et élaborés pouvant accélérer le lent travail de transcription des langues.

3.2. Faciliter la diffusion de connaissances linguistiques

Ainsi, les connaissances linguistiques peuvent être présentées sous forme d'hypertextes téléchargeables à partir des sites des institutions, mais produire ces hypertextes représente une somme de travail qu'il ne semble pas possible de mettre en œuvre à court terme. Il semble plus judicieux de créer des correcteurs orthographiques au sein des éditeurs de textes car il est possible de réaliser ces correcteurs avec peu de travail humain mais en s'appuyant sur des corpus de textes.

Par ailleurs, même si les langues africaines sont actuellement peu présentes sur la Toile [Diki-Kidiri 2003], la levée des difficultés rédhibitoires que constituaient la représentation des caractères spéciaux et la mise au point d'outils d'édition adaptés, devraient permettre la création de sites internet rédigés dans les langues africains pour les locuteurs de ces langues.

3.3. Faciliter la production de textes

Les utilisateurs peuvent également être sollicités pour fournir leurs productions aux institutions afin de collaborer à la constitution des corpus. Cette collaboration suppose que les textes produits soient codés selon le standard Unicode encore mal connu en Afrique hors de la communauté des informaticiens. Elle ne peut se mettre en place que grâce à la mise au point d'outils adaptés, ce qui n'est pas le cas actuellement.

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 ou deux touches, 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 puisque la tâche de saisie s'effectue dans des conditions ergonomiques satisfaisantes et que le texte saisi est directement encodé selon le standard Unicode. Malheureusement le choix de l'emplacement des caractères spéciaux sur le clavier n'est régi par aucun principe général, le risque de prolifération de claviers concurrents est donc réel. Il est donc urgent de définir des règles d'attribution des touches de clavier aux nouveaux caractères.

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 : l'utilisateur doit 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. Pour l'instant, il n'existe pas d'éditeurs développés pour les langues moins dominantes, comme les langues africaines. Toutefois, 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.

4. Objectifs

Nous nous focalisons sur la mise au point de correcteurs orthographiques réalisés à partir de corpus de textes et d'outils permettant à des linguistes de dépouiller facilement des corpus de textes afin d'enrichir les ressources linguistiques.

4.1. Correcteurs orthographiques

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 :

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

1. La détection des erreurs s'effectue 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 travaux les plus récents s'inscrivent dans le domaine de la cohésion lexicale. Mais 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].

2. 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, d'une clé de similarité, ou encore mesure de la distance phonologique).

3. 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).

4. 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.

Des besoins spécifiques

Il existe déjà des correcteurs orthographiques pour certaines langues africaines (Microsoft a annoncé en 2004 la mise sur le marché d'un correcteur orthographique pour Word en Kiswahili), 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.

Enfin, 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.* soutien à la production de ressources linguistiques).

Ce correcteur orthographique est, dans une certaine mesure, indépendant de la langue puisque les traitements dépendants de la langue (comme le calcul des flexions et dérivations des mots par exemple) sont décrits dans des modules génériques qui utilisent les informations 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. 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.

Ce correcteur orthographique 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 a été achevé en 2004. Les tests réalisés début 2005 déboucheront sur la réalisation d'une nouvelle version en 2006.

4.2. 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 capitalisation, 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.

L'enrichissement du lexique au sein de l'institution en charge de la langue peut être facilité par le développement d'un logiciel adéquat intégrant le dépouillement de corpus. 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.),
- concordancier.

Les personnes chargées de la maintenance des ressources électroniques verraient leur tâche facilitée par l'utilisation d'un logiciel et de noter de nouvelles informations (telles la catégorie grammaticale, une définition, etc.) dans des formulaires adaptés. De telles informations peuvent même être suggérées par le logiciel à la suite d'une phase d'apprentissage de la langue à partir des corpus. Ces informations peuvent être directement inscrites dans le lexique électronique de la langue.

4.3. Constitution de ressources linguistiques

Nous avons pour projet de développer des correcteurs orthographiques pour plusieurs langues africaines (dans un premier temps : bambara, kanuri, tamajaq, fulfulde, wolof, hausa et zarma). Nous avons recueilli des ressources textuelles 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 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 2 : corpus recueillis

La constitution de ressources linguistiques électroniques à partir de ces corpus bruts se déroule en plusieurs étapes : Il faut commencer par normaliser les textes qui, bien qu'écrits en langue correcte, respectant les décrets en vigueur, 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. Nous avons ensuite balisé les textes avec des balises XML[11] conformément la norme XCES[12].

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).

5. 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 (développement effectif de correcteurs orthographiques, constitution et diffusion de ressources linguistiques électroniques) , permettant ainsi aux africains d'utiliser leurs propres langues dans les environnements de communication les plus modernes.

La capitalisation de textes électroniques sera également un facteur de progrès et 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.

11 eXtended Markup Language http://www.w3.org/XML/

12 Corpus Encoding Standard for XML http://www.xml-ces.org/

6. Bibliographie

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

[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.

[Dalby 1996] Dalby, D., "L'Afrique et la lettre", Centre Culturel Français, Lagos & Fête de la Lettre, Paris, 1986.

[Diki-Kidiri 2003] Diki-Kidiri, M., Atibakwa Baboya E. (2003) "Les langues africaines sur la toile", Les cahiers du RIFAL n°23 Le traitement informatique des langues africaines, pp. 5-32, novembre 2003.

[Dumont 1983] Dumont, P., "Le français et les langues africaines au Sénégal", Paris, Acct-Karthala, 1983.

[Enguehard 2003] Enguehard, C., Mbodj, C., "Recueillir et diffuser les noms des plantes dans les langues africaines". Les cahiers du RIFAL n°23 Le traitement informatique des langues africaines, pp.47-54, novembre 2003.

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

[Gaucher 1968] Gaucher, J., "Les Débuts de l'enseignement en Afrique francophone. Jean Dard et l'école mutuelle de Saint-Louis du Sénégal" , Paris, Le livre africain, 198 pages, 1968.

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

[Hirst 1998] Hirst, G., St-Onge, D., "Lexical chains as representations of context for the detection and correction of malapropisms", Fellbaum, pp.305-332, 1998.
http://www.cs.toronto.edu/compling/Publications/Abstracts/Papers/Hirst+Budanitsky-2001-abs.html.

[Hirst 2003] Hirst, G., Budanitsky, A., "Correcting real-word spelling errors by restoring lexical cohesion", 2003.

[Mays 1991] Mays, E., Damerau, F. J., Mercer, R. L., "Context based spelling correction", *Information Processing and Management*, 27(5), pp.517-522, 1991.

[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.

Chantal Enguehard est Maître de Conférences en informatique depuis 1993. elle travaille dans le domaine du Traitement Automatique de la Langue Naturelle et s'est spécialisée dans le domaine des langues bénéficiant de peu de ressources linguistiques, qu'il s'agisse de langues non internationales comme les langues africaines, ou de langues internationales dégradées (comme les prises de notes ou les textes de petites annonces).

Depuis 1995, Chantal se rend régulièrement dans la région du Sahel, en particulier au Sénégal, au Niger et au Mali où elle collabore avec les institutions nationales.