Traitement informatique des langues africaines : problèmes et perspectives

Émile Camara
Groupe pour le Traitement Informatiques des Langues dans l'Enseignement
École Normale Supérieure
Bamako, Mali

Josué Ndamba & Célestin Nstadi
Groupe de Recherche Langue et Informatique
Université Marien Ngouabi
et Institut National de Recherche et d'Action Pédagogiques
Brazzaville, Congo

Véronique Rey & Jean Véronis
Laboratoire Parole et Langage, Université de Provence & CNRS
29, Avenue Robert Schuman
13621 Aix-en-Provence Cedex 1, France
Veronique.Rey@lpl.univ-aix.fr, Jean.Veronis@lpl.univ-aix.fr

Mots-clés: traitement informatique des langues, langues africaines.

Résumé: Cet article décrit quelques problèmes et perspectives liés au traitement informatique des langues africaines. Il décrit tout d'abord le contexte technologique, social et culturel dans lequel ce traitement s'insère, puis examine une série de questions techniques. L'article décrit enfin un certain nombre de résultats et de directions pour des développements futurs, à travers l'exemple d'un projet en cours associant des partenaires français, maliens et congolais.

Abstract: This paper outlines a number of problems and perspectives concerning the computerized treatment of African languages. The paper first discusses the technological, social and cultural context of this treatment, and then describes a series of technical aspects. Finally, a number of results and directions for future development are outlined, using the example of a continuing collaborative project among partners in France, Mali and Congo.

1. Introduction

Il est devenu banal de dire que l'informatisation est en train de produire sur les langues une révolution comparable à celles, en leurs temps, de l'imprimerie ou du passage du rouleau de lecture au livre calligraphié. Toutes les opérations de manipulation de documents textuels (rédaction, production, édition et exploitation) sont désormais informatisées. Le réseau Internet parachève cette informatisation, en permettant la diffusion et l'utilisation des documents sous forme purement électronique. Les outils de traitement automatique des langues, à savoir aides à la rédaction (correcteurs orthographiques et grammaticaux), aides à la traduction (dictionnaires en ligne, mémoire de traduction, traducteurs semi-automatiques), aides à la manipulation de documents (recherche documentaire, filtrage et routage automatique de messages), etc., sont amenés à devenir des éléments incontournables de la «société de l'information» naissante.

Le but de cet article est de discuter du développement des technologies et du nécessaire traitement informatique des langues africaines. Nous décrirons tout d'abord le contexte social, technologique et culturel dans lequel ce développement s'insère (section 2), puis nous examinerons les problèmes que pose, du point de vue technique, le traitement informatique des langues africaines (section 3), et nous tracerons enfin un certain nombre de perspectives, à travers l'exemple d'un projet en cours associant des partenaires français, maliens et congolais (section 4).

2. Contexte technique et culturel

2.1. Contexte socio-technologique

L'intérêt potentiel de l'Internet pour les pays africains a été analysé en détail par (entre autres) Bellman et Tindimubona (1991), Sadowsky (1993) et Hills (1993):

- les **listes de discussion électronique** permettent de remplacer en partie les colloques et conférences, qui sont coûteux et donc d'accès difficile;
- les **sites WWW** offrent des masses d'informations et de documents gratuits et perpétuellement mis à jour sur tous les domaines;
- le **courrier électronique** offre des possibilités d'interaction immédiate entre chercheurs, et donc de transfert d'expertise Nord-Sud et d'intégration régionale.

L'Internet pourrait ainsi répondre à un des problèmes majeurs qui se posent aux pays africains en matière d'éducation supérieure: de nombreux étudiants acquièrent à l'heure actuelle un savoir et une expertise dans les pays du Nord, mais se trouvent souvent en position d'isolement lors de leur retour, dans l'incapacité de suivre l'évolution de leur domaine et de communiquer efficacement avec les autres chercheurs au niveau international.

Djamen *et al.* (1995) font remarquer que l'Internet pourrait aussi jouer un rôle symétrique, à savoir faire connaître au reste du monde les productions et la culture africaine. Sur le plan scientifique, par exemple, les chercheurs africains pourront plus facilement communiquer leurs travaux et co-développer des recherches avec les chercheurs des pays industrialisés. Sur le plan littéraire et culturel, l'Internet peut contribuer à faire mieux connaître les pays africains au reste du monde. Des serveurs WWW et des listes de discussion sur l'Afrique sont déjà en activité dans les pays industrialisés et rencontrent un net succès (par exemple, *The African Studies World-Wide Web*[1] de l'University of Pennsilvania a reçu plus de 80000 accès au cours du seul mois de février 1995: voir Sisskind, 1995). Il est vivement souhaitable que des sites de ce type se développent en Afrique même.

Sur le plan technique, la couverture du continent africain par des réseaux informatiques est infiniment plus faible que celle du reste du monde, mais elle est en progression constante. Une trentaine de pays africains ont à l'heure actuelle accès à l'Internet. Une douzaine ont un accès incluant des facilités interactives telles que *telnet* et *ftp*, les autres ayant accès au seul courrier électronique (Crépin-Leblond, 1995). Certains sous-réseaux sont opérationnels depuis plusieurs années, comme celui du RIO, «réseau intertropical d'ordinateurs» de l'ORSTOM[2], qui a 100 points d'accès et 1000 utilisateurs en Afrique. De nombreux autres sous-réseaux sont en développement ou en projet, soit dans l'Internet, soit dans BitNet, FidoNet et UUCP (voir la liste établie par Godard, 1995). Si les projets initiaux ont été gérés par des opérateurs tels que l'ORSTOM, des sous-réseaux Internet nationaux émergent à présent dans de nombreux pays (Sénégal, Mali, Burkina-Faso, etc.) et des écoles d'ingénieurs, établissements de recherche et entreprises privées sont en train d'acquérir la technologie Internet et de prendre le relais.

Il est donc prédictible que les réseaux informatiques vont continuer à se développer en Afrique, et permettront au continent de ne pas rester à l'écart des «autoroutes de l'information», suivant en cela le modèle de l'Amérique latine, par exemple (Primienta, 1992), dont le taux de couverture en progression extrêmement rapide montre bien l'intérêt des réseaux informatiques pour les pays en développement.

2.2. Contexte socio-linguistique

Actuellement, de nombreux pays africains gèrent des politiques linguistiques organisées autour de deux types de langues:

- une **langue officielle**, qui est souvent le français ou l'anglais, mais parfois aussi une langue africaine comme le swahili (Kenya, Tanzanie) ou l'amharique (Éthiopie). La langue officielle permet une communication inter-ethnique, et un accès à une communication internationale, orale et écrite. Elle est en principe enseignée dans le système scolaire.

- des **langues nationales**, langues africaines généralement sélectionnées sur des critères démographiques. Les langues nationales servent souvent à l'alphabétisation et permettent la diffusion d'informations concernant par exemple la prévention dans le domaine médical (Chaudenson et Slozian, 1994), ou la formation professionnelle (mécanique, etc.).

Les langues nationales sont une réalité incontournable: il y a plus de 100 langues nationales différentes pour 52 pays africains (Mazrui, 1986) et elles sont parlées (et parfois écrites) par des millions de locuteurs. Par exemple, le bambara, langue principale du Mali, de la famille des langues manding, est utilisé par environ 3 millions de locuteurs au Mali même, et un million dans les pays voisins (Gambie, Burkina Faso, Sénégal et Côte d'Ivoire: Grimes, 1992).

Par ailleurs, les pays du Nord connaissent des migrations africaines depuis de nombreuses années au point de constituer de véritables diasporas, comme par exemple, en France, les Soninké, originaire de la région du fleuve Sénégal (Quiminal, 1991). De plus, on enseigne de nombreuses langues africaines en Europe et aux États-Unis. Par exemple, en France, l'INALCO (Institut national des langues et civilisations orientales) a cette vocation. Enfin des ressortissants africains produisent en France et dans d'autres pays une cinématographie et une littérature africaine non négligeables, dont *Kaïdara, récit initiatique peul*, (Hampâté Bâ, 1969) est un des nombreux exemples.

La couverture de l'Afrique par l'Internet pourrait soutenir et renforcer cette dynamique linguistique en Afrique et entre l'Afrique et les pays industrialisés.

2.3 Le traitement informatique des langues

Malgré les aspects positifs décrits jusqu'à présent, l'Internet risque d'accroître le phénomène d'appauvrissement linguistique en favorisant l'usage exclusif de l'anglais[3]. La seule façon de contrecarrer cette tendance est de fournir aux langues minoritaires un support informatique adéquat, depuis la simple possibilité de transmission des caractères à travers les réseaux, jusqu'aux outils évolués tels que:

- **aides à l'écriture et à la publication** (dictionnaires en ligne, correction orthographique et grammaticale, etc.);
- **aide à la traduction** (constitution automatique de lexiques et bases terminologiques bilingues, corpus bilingues alignés et annotés, mémoires de traduction, etc.);
- **assistance à la création terminologique** pour les concepts nouveaux (cf. Ntsadi, 1990);
- **enseignement des langues assisté par ordinateur**, qui pourrait jouer un rôle considérable dans le futur, en particulier pour l'aide au bilinguisme (Camara, 1992).
- **aide au feuilletage** ("browsing") et traduction automatique sommaire.

Les langues africaines ont cependant, dans ces domaines, accumulé un retard considérable, qu'il sera difficile de combler sans des actions incitatives de grande ampleur.

La création d'outils informatiques pour les langues africaines pourra également avoir un rôle important dans les politiques linguistiques en donnant aux commissions d'experts des moyens pour poursuivre leurs travaux commencés dès les années 60:

- homogénéisation des données,
- validation des règles d'écriture sur un corpus important,
- dialogue avec d'autres pays africains ou non africains.

3. Caractéristiques des langues africaines

Cette section est destinée à illustrer certaines caractéristiques des langues africaines du point de vue du traitement automatique. Nous explorerons les problèmes liés à la reconnaissance des unités lexicales (depuis les systèmes d'écriture et l'orthographe jusqu'à la morphologie), qui sont la première étape d'à peu près tout traitement informatique. Les langues africaines diffèrent substantiellement de ce point de vue des langues habituellement utilisées en traitement automatique (essentiellement l'anglais et certaines langues européennes, ainsi que des langues

asiatiques telles que le japonais). Les étages ultérieurs du traitement (syntaxe, sémantique) montrent aussi des différences marquées, mais feront l'objet d'une étude ultérieure.

Les caractéristiques énumérées ci-après se traduisent souvent en difficultés du point de vue de la réalisation des programmes informatiques. Mais ces difficultés ne sont nullement insurmontables, et elles sont en même temps extrêmement intéressantes pour le chercheur par leur répercussion sur le traitement des langues en général, y compris des langues européennes. En effet, le traitement informatique des langues a développé au fil des années des modèles euro-centristes--pour ne pas dire anglo-centristes, et il est clair que ces modèles plafonnent à un certain niveau de compétence: on ne dispose toujours pas, malgré quelque quarante ans d'efforts d'une grammaire formelle couvrant de façon large une quelconque langue, y compris l'anglais. Il y a fort à parier que l'examen des langues africaines (du point de vue de la formalisation et de l'automatisation) conduira à une révision des théories et modèles en vogue. Ces langues sont en effet extrêmement différentes des langues (ouest-)européennes, et amènent de nouvelles contraintes dont il faudra tenir compte, ce qui permettra vraisemblablement de gagner en généralité. Il est intéressant de remarquer que certains problèmes sont d'un point de vue informatique analogues à ceux que posent les langues de l'est de l'Europe (slaves et finno-ougriennes en particulier).

3.1 Alphabets

On constate différentes situations selon que les langues ont une ancienne tradition écrite (par exemple, l'amharique en Éthiopie), ou une tradition plus récente basée sur l'alphabet arabe ou latin, complétés avec des caractères de l'Alphabet Phonétique International (API) qui permet des transcriptions de sons spécifiques aux langues africaines (comme les préglottalisées [b], [d], [g]). De plus, de nombreuses langues ont une organisation de type accentuel ou tonal qui mettent en jeu des propriétés prosodiques attachées aux unités significatives élémentaires (Creissels, 1989:174). Ces propriétés sont reflétées dans l'écriture par des signes diacritiques qui se greffent aux caractères latins ou aux symboles de l'API, et dont on ne peut faire l'économie sous peine d'incompréhension. Ainsi, le ngbaka (langue de la famille Niger-Congo, parlée au Zaïre et en RCA) a trois tons ponctuels (haut, moyen, bas), et quatre tons mélodiques (descendant-haut, descendant-bas, montant-haut, montant-bas; Thomas, 1981:209):

mà "magie", mā "pluie", má "comment",
máā "pouvoir", máà "à moi", màá "je", màa "moi".

La variété des jeux de caractères n'est pas sans poser problème aux systèmes informatiques. Aussi trivial que le simple affichage de caractères puisse paraître, il est actuellement résolu de façon imparfaite, comme s'en rend compte tout utilisateur qui transfère un texte français sur les réseaux ou entre deux micro-ordinateurs de marques différentes. La situation est cependant en voie d'amélioration:

- Des normes ont été adoptées sur l'Internet, qui permettent de transporter les 8 bits des caractères sans corruption dans le protocole TCP/IP (par exemple dans les applications *telnet* et *ftp*). De plus la norme MIME (*Multi-purpose Internet Mail Extensions*: RFC-1521 and RFC-1522) permet l'échange de données sans corruption par compactage et décompactage appropriés.
- La série ISO 8859 (qui est un sur-ensemble de l'ASCII ou ISO 646-IRV) permet de représenter les caractères de nombreuses langues à écriture latine, cyrillique, arabe, grecque et hébreu. Ces jeux de caractères ne sont pas implémentés sur tous les ordinateurs, mais ils commencent à être utilisés comme standard sur l'Internet (en particulier ISO 8859-1 pour les langues de l'Europe de l'Ouest).
- Le standard récent ISO 10646-1 (*Universal multiple-octet coded character set* ou UCS) a pour but de fournir un jeu de caractères unifiés pour l'ensemble des langues. Son encodage sur 32 bits permet en théorie des milliards de caractères et seule la première partie (encodable sur 16 bits) a été publiée, et couvre de très nombreuses langues (en particulier asiatiques). Il faudra certainement un temps non négligeable avant que cette norme soit universellement implantée, mais il est probable qu'elle sera la norme du futur.

Les langues africaines sont généralement absentes des débats sur les standards. Ainsi, nombreuses sont les langues africaines qui ne peuvent pas être représentées par la norme ISO 8859 à cause de

l'absence des signes de l'API et des caractères pourvu des diacritiques représentant les tons. Pire, le nouveau standard UCS s'est développé sans inventaire précis des besoins des langues africaines sauf l'amharique. Certes, l'UCS comprend les caractères de l'API, mais rien ne dit que toutes les combinaisons de diacritiques nécessaires y sont présentes.

Enfin, certaines langues peuvent se transcrire dans plusieurs systèmes d'écriture. Par exemple, en wolof (langue nationale du Sénégal, groupe Ouest-Atlantique), l'écriture arabe permet la transcription de textes religieux, mais des règles d'écriture en lettres latines augmentées de caractères API (officialisées en 1975) permettent une autre transcription. Dans ce dernier cas, il peut arriver, que des lecteurs wolofs éprouvent des difficultés à comprendre un texte faute de savoir interpréter les caractères API, ce qui pousse certains auteurs à utiliser *kh* et *gn* au lieu des caractères API (x et ñ). On a ainsi trois systèmes de transcription concurrents pour la même langue, ce qui complique fortement les procédures de traitement automatique. On pourrait noter d'autres cas tels que le berbère (alphabet tifinagh de gauche à doite ou de droite à gauche, ou alphabet arabe), ou le comorien (écriture latine, ou arabe pourvue de diacritiques spéciaux) (Haralambous et Plaice, 1995).

3.2 Orthographe

Beaucoup de langues nationales sont des langues à traditions orales qui ont connu et connaissent des conventions d'écriture fluctuantes. La graphie n'est donc pas toujours fixée et une langue peut avoir plusieurs orthographes possibles. Le Wolof, par exemple, connaît de nombreuses divergences régionales. Certains noms possèdent un *a* final pouvant se noter *ë* dans certaines régions et même disparaître; certains auteurs transcrivent des consonnes doubles, ou font apparaître un *é* et un *à* différent de *e* et *a* pour mentionner l'ouverture de la voyelle.

La segmentation des mots, quant à elle, n'est pas toujours stable: par exemple, des verbes conjugués à des temps composés risquent d'être transcrits sans segmentation (sans blanc typographique) ou bien avec. Creissels (1991:33) indique que les conventions de découpage en mots sont la plupart du temps un simple calque de ce qui existe dans les langues d'Europe. Il donne l'exemple suivant: en bambara, on écrit *misiw* «les vaches»; *a ka bon* «il est gros». La marque du pluriel orthographié *w* est collée au substantif, dont elle est pourtant séparable dans cette langue; le morphème d'affirmation *ka* est séparé du morphème suivant, dont il est rigoureusement inséparable. L'auteur conclut: «une graphie dégagée de l'influence des orthographes européennes aurait certainement abouti au découpage *misi u* "les vaches" et *a kabon* "il est gros"».

Ces phénomènes sont extrêmement difficiles pour les systèmes de traitement automatique. Ils nécessitent la mise en place de mécanisme de recherche flexible, permettant de retrouver les graphies variantes en fonction de règles[4].

Depuis les années 60, une tradition africaine existe en matière de normalisation de la transcription des langues. Par exemple:

- Un comité de standardisation a été formé à l'époque coloniale pour le swahili, qui s'emploie à tous les niveaux de la vie politique et administrative, de l'enseignement et des média (radiodiffusions nationales et internationales: Grande-Bretagne, États-Unis, Belgique, Allemagne, etc.). Actuellement un service spécial au ministère de l'Éducation nationale (Tanzanie et Kenya) et des associations privées s'occupent de la normalisation.
- Dans d'autres pays, des commissions tentent d'établir des critères linguistiques pour définir des conventions d'écriture et les diffuser. Grâce à ces équipes, les langues nationales occupent des espaces culturels. Par exemple au Bénin, la radio nationale diffuse dans 18 langues; des linguistes animent des émissions linguistiques; des concours de dictées sont organisés dans ces langues, et dans les villages, les enfants participent à ces concours, puis, envoient leurs copies à Cotonou pour correction.
- Au Mali, le bambara possède désormais une orthographe officielle[5]. L'orthographe a été officialisée aux fins de l'alphabétisation fonctionnelle par un décret en 1967, fixant l'alphabet pour la transcription des langues nationales. Ce décret fait suite à une réunion d'experts en linguistique et en alphabétisation avec l'aide de l'Unesco en mars 1966.

La pratique écrite aujourd'hui par les usagers nous semble primordiale car elle est déjà une histoire de la langue. Ainsi pour chaque langue le souci sera de connaître les méthodes de transcriptions officielles. On aura ainsi un inventaire des langues écrites disponibles aujourd'hui. L'informatisation des langues, par la création de bases de données et l'inventaire systématique pourra permettre de contribuer à l'effort de normalisation, et à sa cohérence entre langues et dialectes voisins.

3.3 Phonologie

Les caractéristiques phonologiques de nombreuses langues africaines posent un problème d'accès au lexique. En effet, la forme des mots peut varier de façon importante selon le contexte phonologique, et la reconnaissance des unités lexicales ne peut alors se faire de façon réaliste par un simple examen de listes lexicales comme c'est souvent le cas pour le français ou l'anglais. Des algorithmes plus complexes de reconnaissance des formes lexicales doivent être mise en oeuvre.

Un exemple très caractéristique illustrera cette situation: celui des *alternances consonantiques.* Dans beaucoup de familles de langues africaines, les bases nominales et verbales présentent une initiale consonantique qui connaît une réalisation différente en contexte intervocalique. Par exemple, en mwali, (langue de Mohéli, île des Comores), l'initiale consonantique /k/ se réalise /h/ en position intervocalique[6]:

kombe "un coquillage" - mahombe "des coquillages".

En agni, (langue du Sud-Est de la Côte d'Ivoire, le long de la frontière du Ghana (Creissel, 1989), l'alternance est plus complexe:

ɔ̀kà "il reste" - ɔ̀à-há "il est resté"- ɔ̀ŋ-gà "il ne reste pas"

Le cas des *alternances tonales* illustre aussi cela: des formes nominales ont des propriétés tonales qui se manifestent différemment selon les contextes. Ces formes incluent des tons flottants qui se produisent lors de contraction vocalique: le ton associé au deuxième segment devient flottant. Il donne alors lieu soit à des règles de rattachement avec le ton subséquent, soit à des règles d'effacement après avoir éventuellement contribué à modifier les tons voisins. Ces cas se rencontrent dans des cas très simples comme par exemple le substantif suivi d'un adjectif: ce substantif n'aura pas la même forme tonale que dans sa forme isolée. Par exemple, en mende (Niger-Congo, famille des langues mandé, Sierra-Leone) on a (Creissels, 1989:224):

pélè "route" - pèlè í "la route" - pèlé ngà "des routes"
fɛ̌ "pot" - fɛ̀ í "le pot" - fɛ̌ ngà "des pots"

Il faut donc, ici encore élaborer des algorithmes qui permettent de prévoir les réalisations tonales. On peut noter de ce point de vue le travail de Jouannet (1987) sur la modélisation des tons en kinyarwanda.

3.4 Morphosyntaxe

Au niveau morphosyntaxique, les systèmes automatiques de traitement des langues doivent effectuer une opération de base qui consiste à accéder à un lexique pour en ramener la forme de base (lemme) et des informations associées (partie du discours et valeurs d'attributs tels que genre, nombre, etc.). Pour le français ou l'anglais, il existe une assez bonne correspondance entre les mots et les morphèmes, qui, *grosso modo,* autorise l'emploi de méthodes frustes telles que l'examen de listes de formes fléchies. Ainsi, à *chiens* sera associée l'information complexe du type *[base=chien, catégorie=nom, genre=masculin, nombre=pluriel]*. Certes, le système verbal du français demande l'enregistrement d'un nombre important de formes, mais la capacité actuelle de stockage des machines ne rend pas absolument flagrant l'intérêt d'une analyse des formes par des règles de décomposition.

Par contre, les langues africaines sont souvent d'un type agglutinatif et la combinatoire des morphèmes produirait des listes de mots trop importantes. Ainsi, le syntagme verbal des langues bantoues est construit à partir d'un radical sur lequel se greffent de nombreuses affixes, dérivations et marqueurs temporels. En mwali, par exemple, on trouve des combinaisons telles que

ngarimhuliaolo, "nous le lui achetons"

qui se décompose de la façon suivante:

nga	indice
ni	nous
m	lui
hul	acheter
i	dérivation verbale "acheter à"
a	marque temporelle
o	marque temporelle
lo	le

Ce phénomène est d'une certaine façon semblable à ce que l'on observe en italien, où certains clitiques peuvent s'agglutiner à la forme verbale (ex.: *damelo* = *da* + *me* + *lo*, "donne-moi le"), mais la combinatoire est d'une productivité et d'une complexité qui ne permettent pas d'une façon réaliste le traitement de la morphologie par simple examen de listes lexicales. On doit nécessairement faire appel à des algorithmes morphologiques complexes.

Par ailleurs, les catégories grammaticales elles-mêmes, parties du discours ou attributs morphologiques, sont souvent éloignées de ce que l'on observe dans les langues européennes. Par exemple, dans les langues bantoues, les substantifs ont un préfixe et un thème et sont organisés en "classes", numérotées par les linguistes de 1 à 15, qui régissent le préfixe à employer avec l'adjectif subséquent, et le substitutif à employer si nécessaire.

Par exemple, en mwali:

muri "arbre" - muri m(u)kundu "un arbre rouge"
mawe "pierres"- mawe makundu "des pierres rouges"

Certains préfixes, les locatifs, ont de plus un statut particulier en donnant différents sens de lieu sur un même thème substantival. Ces morphèmes qui servent à former des locatifs ne sont ni des prépositions, ni des propositions, mais de simples préfixes. De plus, ils permettent aux substantifs sur lesquels ils se greffent d'occuper la position sujet: il y a donc des phrases à sujet locatif dans les langues bantoues.

Comme on le voit, le fonctionnement morphosyntaxique des langues africaines est fortement différent de celui des langues européennes, et il est à prévoir que les méthodes et algorithmes de traitement doivent être révisés. Par ailleurs, les efforts d'harmonisation et de standardisation en cours au sein de projets internationaux tels que MULTEXT et EAGLES (Véronis et Khouri, 1995) ont pour l'instant une vision très eurocentriste des phénomènes linguistiques et il est certain que la prise en compte de langues africaines obligera à remettre en cause bon nombre de schémas.

4. Premiers résultats et perspectives

A travers une Action de Recherche Partagée (ARP), récemment mise en place grâce à un financement de l'AUPELF•UREF (ARP ALAF: *Alignement des Langues Africaines et du Français*), nous espérons montrer qu'un partenariat entre instituts du Nord et du Sud peut permettre une informatisation des langues africaines avec des ressources limitées. L'idée générale de cette action est de faire bénéficier les langues africaines des efforts réalisés dans des projets internationaux tels que les Actions de Recherche Concertée (ARC) de l'AUPELF•UREF, MULTEXT (Ide et Véronis, 1994)[7], EAGLES ou la *Text Encoding Initiative* (Ide et Véronis, 1995) et qui portent sur le développement de standards, de méthodologies et d'outils multilingues.

L'ARP ALAF est un partenariat entre le Laboratoire Parole et Langage, l'Ecole Normale Supérieure du Mali, l'Université Marien Ngouabi et l'Institut National de Recherche et d'Action Pédagogiques au Congo dont le but est:

- de créer un ensemble minimal de ressources informatisées (lexiques et textes bilingues avec le français pour langue cible);

- de transférer aux équipes du Sud une expertise concernant les techniques récentes d'ingénierie linguistique;
- et plus généralement, d'établir et de maintenir des réseaux de relations entre universités du Nord et du Sud, dans la conjoncture actuelle.

L'ARP porte sur trois langues: la bambara (parlé au Mali), le kikongo (parlé au Congo) et le swahili (en raison de son rôle géo-politique). Pour chacune de ces langues, nous avons commencé à constituer des corpus informatisés, soit à partir de textes scannés, soit à partir de données informatisées pré-existantes et converties dans un format standardisé (SGML). Un texte d'environ 130000 mots, le Nouveau Testament, existe dans les trois langues, et permettra une étude comparative.

Des lexiques sont également en cours de réalisation, contenant pour l'instant 2090 formes pour le bambara, 5075 pour le kikongo, et 27758 pour le swahili (où il a été possible d'utiliser comme point de départ le lexique rélisé dans le projet *Kamusi*[8] (Yale University). Un ensemble de catégories morphosyntaxiques a été défini pour les trois langues. En ce qui concerne le système nominal, le choix s'est porté sur la saisie de toutes les formes, étant donné la combinatoire relativement réduite. Ainsi par exemple, en kikongo et en swahili, un adjectif apparaîtra avec l'ensemble des accords des classes nominales. De même, chaque nom commun apparaît sous sa forme au singulier et celle du pluriel. En bambara, la choix est d'autant plus justifié qu'il n'y a pas de flexion sur le thème nominal. Par contre, les systèmes verbaux, en kikongo et en swahili, sont morphologiquement plus complexes. Notre premier choix fut d'extraire les formes du texte, et non de les construire a priori. Ainsi, nous obtenons en swahili 12557 formes verbales sur l'ensemble du Nouveau Testament dont 316 formes pour le verbe *kufanya* "faire" (sur plus de 2000 possibles). Ceci démontre clairement la nécessité d'une analyse morphologique automatique, et nous avons commencé à élaborer des graphes de constructions verbales qui constitueront la base d'un analyseur.

5. Conclusion

La réalisation d'outils informatiques pour les langues africaines semble être un enjeu important pour la place de ces langues dans le monde moderne basé sur les communications, et pour celles des cultures qu'elles représentent. Nous avons examiné dans cet article les différences les plus marquées entre langues africaines et langues européennes du point de vue du traitement automatique. Ces différences constituent sans doutes des difficultés, mais elles sont aussi l'occasion de tester les méthodologies et techniques en cours de développement en Europe, et il est probable que les systèmes y gagneront en efficacité et en généralité. Nous avons mentionné une expérience concrète qui se met en place entre le CNRS et deux équipes africaines, et qui vise à réutiliser des méthodologies, techniques et outils développés dans le cadre multilingue de gros projets de recherches européens, et ainsi à parvenir à des résultats concrets avec un coût de développement limité. Une telle expérience de transfert a déjà été fructueuse pour des langues de l'Europe de l'Est et des langues européennes minoritaires. Si l'expérience en cours avec des partenaires africains est également positive, elle pourra servir de modèle pour de nombreuses autres langues africaines.

Notes

1 <URL: http://www.sas.upenn.edu/African_Studies/AS.html>

2 <URL:http://www.orstom.fr/rio/rio.html>

3 Voir la discussion sur ce sujet dans *Le Monde Diplomatique*, mai 1996.

4 Il est intéressant de remarquer que ce problème se retrouve dans les langues régionales européennes telles que l'Occitan, où la graphie n'est pas complètement fixée, et diffère selon les régions et les dialectes.

5 Il est à noter que la graphie a été fixée mais non les règles relatives à la coupe morphologique des mots dans la phrase.

6 Les exemples mwali sont dûs à Véronique Rey (étude en cours sur les langues comoriennes).

7 <URL: http://www.lpl.univ-aix.fr/projects/multext/>

8 <URL: http://www.yale.edu/swahili/>

Remerciements

Ce travail a bénéficié du concours de l'Agence francophone pour l'enseignement supérieur et la recherche (AUPELF•UREF; convention X14.10.03.01.1/95.04.1) Les auteurs tiennent par ailleurs à remercier Liliane Khouri pour son support technique, ainsi que Jean Doneux pour ses commentaires et son aide précieuse.

Références bibliographiques

Bellman, B. L., Tindimubona, A. (1991). Global networks and international communications: Afrinet. *34th Annual Meeting of the African Studies Association.*

Camara, E. (1992). Les didacticiels des langues: cas des didacticiels du Bambara. L'enseignement assisté par Ordinateur, *Congrés de l'Association Internationale de Pedagogie Universitaire*, Yaounde.

Chaudenson, R., Slozian M. (1994) *Comprendre pour communiquier et soigner: langues, informatique et sajté oculaire en Afrique*, Didier-Erudition, Paris.

Creissel D. (1989). *Aperçu sur les structures phonologiques des langues négro-africaines.* ELLUG, Grenoble.

Creissels, D. (1991). *Description des langues négro-africaines et théorie syntaxique*, ELLUG, Grenoble.

Crépin-Leblond, O.M.J. (1995). *International E-mail Accessibility.* Version du 01/12/95: <URL:http://www.ee.ic.ac.uk/misc/country-codes.html>.

Djamen, J.Y, Ramazani, D., Soteg Some, S. (1995). Networking in Africa: An Unavoidable Evolution towards the Internet. *African Regional Symposium on Telematics for Development*, Addis-Abbeba, Ethiopia, 3-7 April 1995. <URL:http://www.sas.upenn.edu/African_Studies/Padis/telmatics_Djamen.html>

Godard, P. (1995). Africa and Sciences: The Availability of Computer Communications. *African Regional Symposium on Telematics for Development*, Addis-Abbeba, Ethiopia, 3-7 April 1995. <URL:http://www.sas.upenn.edu/African_Studies/Padis/Godard.html>

Grimes, B.F. (Ed.), 1992. *Ethnologue: Languages of the World,* twelfth edition, Dallas: Summer Institute of Linguistics. Disponible sur Internet en version mise à jour (*The Ethnologue Database*). <URL::http://www-ala.doc.ic.ac.uk/~rap/Ethnologue/>

Hampâté Bâ, H. (1969). *Kaïdara, récit initiatique peul.* Paris, 1969. Collection Classiques africains, Les Belles lettres (version poétique bilingue).

Haralambous, Y., Plaice, J. (1995). Ω, une extension de TEX incluant UNICODE et des filtres de type Lex, *Cahiers GUTenberg,* 20, 55-79.

Hills, J. (1993). Telecommunications and democracy: the international experience. *Telecommunication Journal,* 60, I, 21--29.

Ide, N., Véronis, J. (1994). MULTEXT (Multilingual Tools and Corpora). *Proceedings of the 14th International Conference on Computational Linguistics, COLING'94*, Kyoto, Japan, 90-96.

Ide, N., Véronis, J. (Ed.) (1995). The Text Encoding Initiative: background and context. *Kluwer Academic Publishers.* Dordrecht.

Jouannet, F. (Ed.) (1987). *Modèle informatisé du traitement des tons (domaine bantou).* SELAF, Paris. ISBN: 2.85297-204-2.

Ntsadi, C. (1990). Des lexiques nouveaux en langues congolaises pour traduire la science et la technologie modernes. *Commission Nationale Congolaise pour l'UNESCO. Revue d'Information n° 5.* Brazzaville.

Ntsadi, C. (1994). Les noms de nombre en langues congolaises: monument sociolinguistique. *Commission Nationale Congolaise pour l'UNESCO. Revue d'Information n° 8.* Brazzaville.

Lumwamu, F., Missakiri, N., Ntsadi, C., Tirvassen, R. (1993). *Les enfants, les langues et l'école: les cas du Congo et de Maurice.* Coll. Langues et Developpement, Didier-Erudition, Paris.

Mazrui, A.A., (Ed.) (1986). General History of Africa VII, Africa since 1935. *Unesco General History of Africa.* Unesco.

Primienta, D. (1992). Research networks in developing countries: Analysis, methodological principles and guidelines for starting. In INET (1992).

Quiminal, C. (1991). *Gens d'ici, gens d'ailleurs.* C Bourgeois éd., Breteuil-sur-Iton.

Sadowsky, G. (1993) Network connectivity for developing countries. *Communications of the ACM,* 36, 8, 42--47.

Sisskind, J. (1995). The African Studies World-Wide Web: University of Pennsylvania. *African Regional Symposium on Telematics for Development*, Addis-Abbeba, Ethiopia, 3-7 April 1995. <URL:http://www.sas.upenn.edu/African_Studies/Padis/telmatics_Sisskind.html>

Unesco (1992). *CIPSH-UNESCO "Red book" project on Endangered Languages in the World.* <URL:ftp://coombs.anu.edu.au/coombspapers/unesco-endangered-languages-project>.

Thomas, J.M.C. (1981). Les langues dans le monde ancien et moderne, CNRS, Paris.
Véronis, J., Khouri, L. (1995). Etiquetage grammatical multilingue: le projet MULTEXT. *Traitement Automatique des Langues,* 36, 1/2, 233-248.
Woodfield , A., Brickley, D. (1995). *The Conservation of Endangered Languages.* Seminar, April 21st 1995, Bristol University: <http://www.bris.ac.uk/Depts/Philosophy/CTLL/report.txt>

7. Normes et standards

RFC-1521
Borenstein N., Freed, N. (1993). *MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies*, RFC 1521, Bellcore, Innosoft, September 1993. <URL:ftp://ds.internic.net/rfc/rfc1521.txt>
RFC-1522
Moore, K. (1993). *Representation of Non-Ascii Text in Internet Message Headers,* RFC 1522, University of Tennessee, September 1993. <URL:ftp://ds.internic.net/rfc/rfc1522.txt>
ISO 646.IRV:1991
Information Processing -- ISO 7-bit coded character set for information interchange. International Standards Organisation, Geneva, 1991.
ISO-8859 (parts 1-10)
Information Processing -- 8-bit Single-Byte Coded Graphic Character Sets - International Standards Organisation, Geneva, 1987-1992.
ISO/IEC 10646-1:1993
Information technology - Character sets and information coding -Universal multiple-octet coded character set - Part 1 - Architecture and basic multilingual plane. International Standards Organisation, Geneva, 1993.