EXPÉRIMENTATION DE NORMES DE BALISAGE EN LANGUES PARTENAIRES

Thierno Cisse
Chérif Mbodj
Marc Van Campenhoudt
Mohamédoune Wane

1 INTRODUCTION

Le réseau Internet, devenu une réalité accessible, permet aux chercheurs des universités africaines d'envisager une meilleure diffusion de leurs travaux de description linguistique et de s'affirmer comme acteurs autant que comme consommateurs. Jusqu'à présent, leurs données linguistiques (enquêtes de terrain, lexiques, transcriptions phonétiques...) ont été consignées à l'aide de toutes sortes de supports et ne peuvent être consultées que sur place. Dans le même temps, on observe que des universités et des ONG du Nord diffusent des données sur les langues d'Afrique, ce qui conduit, *de facto*, à occulter tout le travail de description et d'appropriation des départements de linguistique des universités africaines. Aujourd'hui, l'adoption de techniques standardisées de consignation des données constitue une étape cruciale pour le développement de l'ingénierie linguistique en langues partenaires ; il importe donc que la communauté des africanistes puisse disposer de normes de description aisément compatibles.

Le langage de balisage XML autorise la représentation d'une grande variété d'informations descriptives au sein d'un corpus linguistique. Grâce au puissant mécanisme des feuilles de style XSL, il permet, dans un second temps, de diffuser les contenus au format HTML utilisé sur Internet. XML présente aussi l'avantage majeur d'avoir été pensé dans la logique du standard Unicode[1], qui autorise l'utilisation d'une grande variété de systèmes d'écriture au sein d'un même document et facilite donc l'usage de caractères issus de l'alphabet phonétique international.

Financée par le réseau *Lexicologie, terminologie, traduction* de l'Agence universitaire de la Francophonie, l'action de recherche en réseau *Expérimentation de normes de balisage en langues partenaires* réunit trois partenaires : l'Université Cheikh Anta Diop de Dakar (Centre de linguistique appliquée de Dakar et département de linguistique), l'Université de Nouakchott (Département des langues nationales et de linguistique) et l'Institut supérieur de traducteurs et interprètes (Haute École de Bruxelles, Centre de recherche Termisti). Elle

1. www.unicode.org/standard/translations/french.html

a pour ambition de permettre aux chercheurs du Nord et du Sud de s'initier au balisage XML des corpus textuels et des bases de données lexicales et de vérifier l'intérêt des propositions de normes existantes au regard de la réalité de langues souvent négligées par la normalisation internationale. L'objectif des partenaires du projet est d'arriver à diffuser, au départ du serveur de leurs universités respectives, des données linguistiques de natures diverses, en sorte qu'elles puissent servir à une large communauté scientifique. La mise en réseau de ces matériaux devrait permettre leur étude à distance et alimenter les travaux des enseignants et des étudiants intéressés par la connaissance de ces langues.

2 XML

Les corpus linguistiques, textuels ou lexicaux, constituent par excellence des documents structurés. Il n'est point nécessaire d'encore en apporter ici l'explication ou la démonstration. L'enjeu fondamental nous semble plutôt être de généraliser la consignation des données linguistiques sous la forme de documents structurés qui soient balisés selon la norme XML[2]. Alors que l'informatique est un outil, l'apprentissage des techniques de balisage peut paraître une corvée fastidieuse là où des suites bureautiques semblent souvent apporter une réponse suffisante aux besoins immédiats. Plutôt que de revenir sur le fait que ces suites ne permettent généralement pas de bien représenter les liens de dépendance qui structurent les données, il nous semble important de souligner l'autonomie nouvelle qu'apporte XML au linguiste. En effet, à l'aide d'un bon éditeur XML, celui-ci peut tout à la fois encoder ses données sous un format universel aisé à exploiter, les diffuser sur un serveur Internet ou Intranet au format HTML et les ordonnancer de la manière qui lui paraît la plus pertinente.

Dans le cas des langues partenaires possédant des caractères phonétiques particuliers, voire une représentation des tons, l'usage de XML semble particulièrement intéressant, puisqu'il intègre un jeu de caractères universel à travers l'usage de la norme Unicode. Les caractères Unicode d'un fichier XML ne sont pas ambigus, ce qui offre la garantie qu'ils seront toujours interprétés correctement. La possibilité d'utiliser aisément Unicode constitue un préalable important qui sera approfondi au point 3.

S'agissant d'expérimenter la pertinence d'un balisage XML, les partenaires de l'action de recherche se sont déjà intéressés à plusieurs définitions de type de document (DTD) proposées comme normes d'échange de corpus linguistiques.

---

2. Les partenaires du projet se doivent de souligner les efforts du Réseau international francophone d'aménagement linguistique (Rifal), qui a pris l'initiative d'offrir à ses partenaires du Sud une première initiation à XML en 2003. Cette formation a été à l'origine de leur intérêt pour cette matière.

2.1 Corpus textuels

Les partenaires se sont principalement intéressés à la norme XCES. Il s'agit d'une proposition d'adaptation à XML de la norme SGML *Corpus Encoding Standard* (CES) qui a résulté des projets européens *Multext* et *Eagles* et est, elle-même, une évolution de la *Text Encoding Initiative* (TEI). Le projet XCES est le fruit d'une collaboration entre le Vassar College (New York) et le Loria (Nancy). Il a débouché sur l'élaboration de feuilles de style permettant de convertir des corpus textuels de XML vers HTML de manière à pouvoir les diffuser sur la toile[3].

Plusieurs essais de balisage XCES de textes en bambara, pulaar et wolof ont été réalisés dans le cadre de l'action de recherche[4]. Quand bien même une bonne maîtrise du formalisme des DTD et de l'anglais est nécessaire pour bien utiliser les balises, force est de constater que celles-ci permettent de rendre compte d'un grand nombre d'éléments descriptifs, même si l'on attend avec intérêt le complément prévu par les concepteurs de XCES pour ce qui concerne les corpus oraux.

La feuille de style *cesdoc.xsl* qui permet de transformer un fichier XCES en fichier HTML est particulièrement puissante. Son expérimentation montre toutefois qu'elle devrait encore être améliorée sous plus d'un aspect pour permettre une présentation qui satisfasse les plus exigeants[5]. On a aussi observé quelques cas où le fichier HTML résultant de la transformation XSL présentait une perte de données[6]. La complexité de cette feuille de style suppose des compétences très avancées de la part de celui qui souhaiterait la modifier, compétences qui ne sont normalement pas celle du linguiste descripteur. La tentation peut être grande de modifier le balisage pour obtenir une présentation HTML adéquate, ce qui serait une mauvaise stratégie. Les partenaires de l'action de recherche plaideraient plutôt, et bien volontiers, pour le financement d'un nouveau développement d'une semblable feuille de style.

2.2 Corpus lexicaux

Beaucoup de lexiques consacrés aux langues nationales ont été réalisés et diffusés avec les moyens du bord et leurs supports informatiques n'ont pas toujours pu être correctement conservés. Ce constat justifie pleinement l'idée d'une consignation sous un format universel. Nous n'approfondirons pas ici la critique des logiciels dédiés à la gestion

---

3. Membre du groupe de travail « formation » du Rifal, Andrei Popescu-Belis (ISSCO, Genève) a le premier attiré l'attention des partenaires de l'action de recherche sur l'intérêt de la norme XCES pour la valorisation des corpus en langues africaines. Ils tiennent à le remercier vivement pour l'aide qu'il a pu régulièrement apporter à leurs travaux.

4. Tous les corpus sont disponibles sur le site du projet : www.termisti.refer.org/ltt/ltt.htm.

5. On songe, par exemple, à la gestion des notes de bas de page ou à la numérotation automatique des titres, difficile à neutraliser.

6. Le phénomène s'est produit lors de la transformation expérimentale du conte wolof anonyme *Doomu Yàlla.*

lexicale ou terminologique : on se bornera à constater que nombre de ces logiciels répondent mal à l'idéal d'universalité (*cf.* 3.1 et 3.2.4) et qu'ils ne permettent généralement pas d'exploiter pleinement la puissance d'un modèle de document (DTD) pour contrôler la succession des champs et l'architecture des données. Aujourd'hui, un logiciel tel que *XML Spy* donne une idée des possibilités offertes par un éditeur XML capable de proposer une grille de saisie comparable à celle d'un gestionnaire de base de données. Une convergence de ces outils peut raisonnablement être attendue au cours des prochaines années.

2.2.1 APPROCHE TERMINOGRAPHIQUE OU LEXICOGRAPHIQUE ?

Les véritables dictionnaires terminologiques adoptent une perspective conceptuelle fondée tout à la fois sur le regroupement des synonymes autour d'une même définition et sur le dégroupement homonymique de termes considérés sous l'angle de la monosémie. Les dictionnaires lexicographiques adoptent plus volontiers un point de vue polysémique et un classement alphabétique. Quelle que soit l'approche initialement suivie, une bonne représentation XML des données devrait permettre par le mécanisme des transformations XSL de passer d'une perspective à l'autre, comme l'a déjà démontré le projet européen DHYDRO (Descotte *et al.* 2001 et Van Campenhoudt 2002).

L'analyse montre qu'une majorité des lexiques conçus dans le cadre de politiques d'aménagement linguistique en Afrique francophone suivent l'approche lexicographique. Conçu très souvent comme une liste alphabétique d'équivalents, ils suivent habituellement une structure simple, mais qui – comme celle de tout dictionnaire classique - n'est pas toujours aussi rigoureuse qu'un modèle de données décrit dans une DTD. La logique voudrait que pour une perspective lexicographique, on utilise les prescriptions du chapitre 12 « *Print Dictionaries* » de la *Text Encoding Initiative* (TEI). Le balisage initialement proposé par cette norme ne permettait pas de décrire véritablement les structures profondes du dictionnaire car il ne prévoyait qu'un codage implicite des informations (Ide et Véronis 1996 : 174). Les partenaires expérimenteront prochainement la récente version P4 de la TEI, réputée compatible avec XML, pour vérifier si elle permet - comme ses auteurs l'affirment - de désormais surmonter cet obstacle (Sperberg-McQueen, et Burnard 2002 : chap. 12.5).

2.2.2 EXPÉRIENCES MENÉES

S'agissant de décrire des données conçues dans une perspective clairement lexicographique, les partenaires ont créé une DTD « maison » pour baliser un échantillon du projet de dictionnaire wolof-français préparé au sein du Département de linguistique de l'Université Cheikh Anta Diop. Les noms des balises sont empruntés à la norme terminologique Iso 12 620, ce qui a l'avantage de permettre une fine description du contenu réel des champs utilisés. L'échantillon balisé a fait l'objet d'une transformation vers HTML consultable sur le site du projet. Dans un avenir prochain, on pourra tenter de représenter également ces données à l'aide de la TEI.

Les partenaires ont également expérimenté une DTD proposée par le Rifal pour baliser l'ensemble des lexiques publiés avec son aide et ensuite les réunir dans une base de données commune. Cette DTD a été conçue dans une perspective conceptuelle et constitue un langage de balisage terminologique (*terminological markup language*) au sens où l'entend la norme Iso 12 642 (2003). Son expérimentation a conduit à ajouter trois éléments supplémentaires (prononciation, classe et renvoi) à cette DTD sans doute trop minimaliste au regard de la réalité des deux échantillons de lexiques balisés[7]. Une première feuille de style a permis une représentation HTML typiquement conceptuelle de ces données[8]. Il sera intéressant de développer ultérieurement une feuille de style permettant de réordonnancer ces mêmes données selon une macrostructure plus lexicographique.

3 UNICODE, L'INDISPENSABLE COMPLÉMENT

L'écriture des langues africaines intègre souvent des caractères ou des signes diacritiques issus de l'alphabet phonétique international. Leur usage a parfois été entériné par la législation. Nombre de travaux de linguistique descriptive et de textes officiels utilisent ces caractères, dont la représentation informatique a longtemps soulevé des problèmes de portabilité.

Le codage des caractères sur 8 bits a, jusqu'il y a peu, limité l'étendue des tables de caractères, contraignant les africanistes à opérer des jeux de substitution arbitraires parmi les différentes tables issues des normes Iso-CEI 8859. Ce pis-aller a débouché sur la mise en circulation de toutes sortes de polices – commerciales ou non – modifiant de manière anarchique les tables Iso-CEI 8859 (1998-2001) et compliquant tout échange de données. Ces anciennes polices satisfont toujours certains besoins locaux, mais elles doivent aujourd'hui être vues comme un obstacle à une bonne représentation informatique des langues partenaires et à une large implantation de leur orthographe. En effet, la norme Unicode permet désormais de représenter en une table unique les caractères d'un très grand nombre de langues, avec pour net avantage qu'un seul et même fichier informatique peut combiner de nombreuses écritures différentes. L'adoption de cette nouvelle norme offre à ses utilisateurs la certitude que leurs textes pourront être lus partout dans le monde avec un affichage correct de leurs caractères, notamment sur Internet.

3.1 Encoder en Unicode à l'aide d'un clavier virtuel

On comprendra aisément que l'adoption d'Unicode – parfaitement adapté à XML - constitue désormais un préalable fondamental pour la bonne représentation et la bonne diffusion des langues partenaires, et ce d'autant plus qu'un transcodage des anciens fichiers utilisant les normes Iso-CEI 8859 est parfaitement réalisable (Chanard et Popescu-Belis

7. *Vocabulaire des élections* (Dialo *et al.* 1997) et *Pour une terminologie de la santé en wolof* (Mbodj 2002).

8. La feuille de style est une adaptation de celle déjà proposée au Rifal par Andrei Popescu-Belis.

2001). L'usage d'Unicode est aisé dès lors que l'on utilise un PC utilisant *Windows 2000* ou *XP*[9]. Il est, bien entendu indispensable d'utiliser une police Unicode (*Arial MS Unicode, Lucida sans Unicode, Sil Doulos Unicode, Gentium,* etc.) pour obtenir un affichage correct des caractères.

Nombre de logiciels sont déclarés « compatibles Unicode ». Malheureusement, l'encodage y est souvent conditionné par le choix d'une langue prévue dans le logiciel et censée utiliser un sous-ensemble limité de caractères Unicode. Lorsqu'il veut encoder des données dans une langue utilisant d'autres caractères, l'utilisateur doit soit changer le clavier actif dans le système d'exploitation (s'il existe), soit utiliser un fastidieux menu d'insertion de caractères spéciaux. Un logiciel réellement conçu dans la logique d'Unicode doit permettre la création de fichiers mélangeant les systèmes d'écriture et un encodage par le biais d'un clavier virtuel, qui réaffecte certaines touches du clavier aux caractères Unicode souhaités.

3.2 Procédure de création d'un clavier virtuel

Les partenaires de l'action de recherche ont créé à titre expérimental des claviers virtuels pour sept langues partenaires de Mauritanie et du Sénégal : le balante, le bambara, le pulaar, le serer et le wolof. Ceux-ci peuvent être téléchargés sur le site Internet du projet (*cf.* note 4).

3.2.1 UN CLAVIER, POUR QUOI FAIRE ?

L'acte même de création d'un clavier doit être mûrement réfléchi en fonction de son contexte d'utilisation : s'agit-il de proposer une norme ou de réaliser un clavier adapté aux besoins d'un chercheur particulier ? Dans le premier cas, il conviendra de tenir compte des législations nationales en matière de représentation des caractères de la langue considérée ainsi que du prescrit des normes internationales (Iso-CEI 9 995-8 1994 et Iso-CEI 14 755 1997). Dans le second, on jugera peut-être utile d'adjoindre une série de caractères phonétiques utiles au travail de description envisagé. Quelle que soit la perspective retenue, on devra viser un objectif d'ergonomie susceptible de faciliter la tâche de l'utilisateur.

Dès lors qu'ils entendaient surtout examiner l'intérêt et la viabilité d'une solution Unicode, les partenaires du projet ne se sont pas souciés de produire des claviers destinés à servir de standards pour les langues considérées. À l'exception du clavier balante, spécifiquement conçu pour les besoins de description d'un chercheur particulier, les autres claviers ont été pensés sur la base des impératifs suivants :

- Disposer les caractères issus de l'alphabet phonétique international à proximité du caractère latin le plus proche : « ɓ » à proximité de « p », « ƭ » près de « t », « ɔ » près de « o », etc.

9. L'expérience montre qu'une simple augmentation à 128 mégaoctets de la mémoire vive d'un ordinateur de type *Pentium III* suffit à permettre l'utilisation de ces systèmes d'exploitation.

- Permettre un usage des caractères propres aux autres langues auxquelles les emprunts sont fréquents (laisser, p.ex., le « q » et le « x » en bambara). Ce choix se heurte toutefois rapidement à la pénurie de touches directement accessibles. Pour conserver les voyelles accentuées du français – mais avec un accès moins direct -, on pourra, par exemple recourir à des combinaisons avec la touche *AltGr.* Il reste à vérifier si dans un tel cas, le simple basculement d'un clavier à l'autre via un raccourci ne se révèle pas aussi efficace.

- Permettre un usage du clavier dans le cadre d'un véritable projet éditorial (auteur, journaliste, enseignant, juriste...) : il convient, notamment, de rendre possible un usage des capitales correspondant à chaque lettre de l'écriture prise en compte.

3.2.2 INVENTAIRE DES CARACTÈRES NÉCESSAIRES

La première démarche consiste à dresser l'inventaire des caractères Unicode qui ne figurent pas sur le clavier français alors qu'ils sont nécessaires pour encoder les textes de la langue considérée. Pour ce faire, on identifiera avec précision les caractères dans les tables Unicode (figure 1). Celles-ci peuvent être visualisées librement sur la toile ou consultées à l'aide d'un petit logiciel dédié[10].

On veillera à décrire chaque caractère en minuscule et en capitale dans un document qui précisera :

- le dessin du caractère ;
- le nom officiel du caractère ;
- le bloc Unicode dans lequel il est présent (le plus souvent : *Latin de base, Supplément latin-1, Latin étendu-A, Latin étendu-B, Alphabet phonétique international (API)* et *Lettres modificatives avec chasse*) ;
- sa notation en Unicode ;
- sa notation sous forme d'entité en XML.

ɲ	Ɲ
nom : lettre minuscule latine N hameçon à gauche	**nom :** lettre majuscule latine N hameçon à gauche
bloc : extensions IPA	**bloc :** latin étendu B
notation Unicode : U+0272	**notation Unicode :** U+019D
entité : ɲ	**entité :** Ɲ

Figure 1 : exemple de descriptif des caractères Unicode nécessaires à la constitution d'un clavier virtuel *Keyman*

10. Le site du projet fournit une liste de sites et de logiciels.

3.2.3 CRÉATION ET DOCUMENTATION DU CLAVIER VIRTUEL

On trouve facilement sur la toile des gratuiciels qui permettent de créer des claviers qui apparaissent sur l'écran. La saisie s'effectue en activant chaque touche à l'aide de la souris, ce qui s'avère rapidement fastidieux. Préférence nous semble devoir plutôt être donnée à un logiciel qui permette de réaffecter chacune des touches du clavier physique, en sorte que l'utilisateur puisse dactylographier son texte immédiatement à l'aide du clavier.

Les partenaires du projet de recherche ont expérimenté le logiciel *Keyman 6*[11] qui permet d'utiliser le clavier physique, tout en faisant apparaître son image en transparence sur l'écran si on le souhaite (figure 2). Le basculement d'un clavier à l'autre s'effectue exactement comme dans le système *Windows,* via la barre des tâches. Ce logiciel d'un coût modique n'existe malheureusement pas en français et présente encore quelques imperfections lors de la création d'un nouveau clavier virtuel. Il permet néanmoins de créer un produit robuste et bien documenté (figure 3). Ce dernier aspect est important, car l'utilisateur final doit pouvoir aisément identifier les touches dont il a besoin. Nous ne reviendrons pas ici sur le détail des améliorations à apporter à ce produit particulier, préférant nous pencher sur la problématique générale de l'usage d'un semblable clavier virtuel.

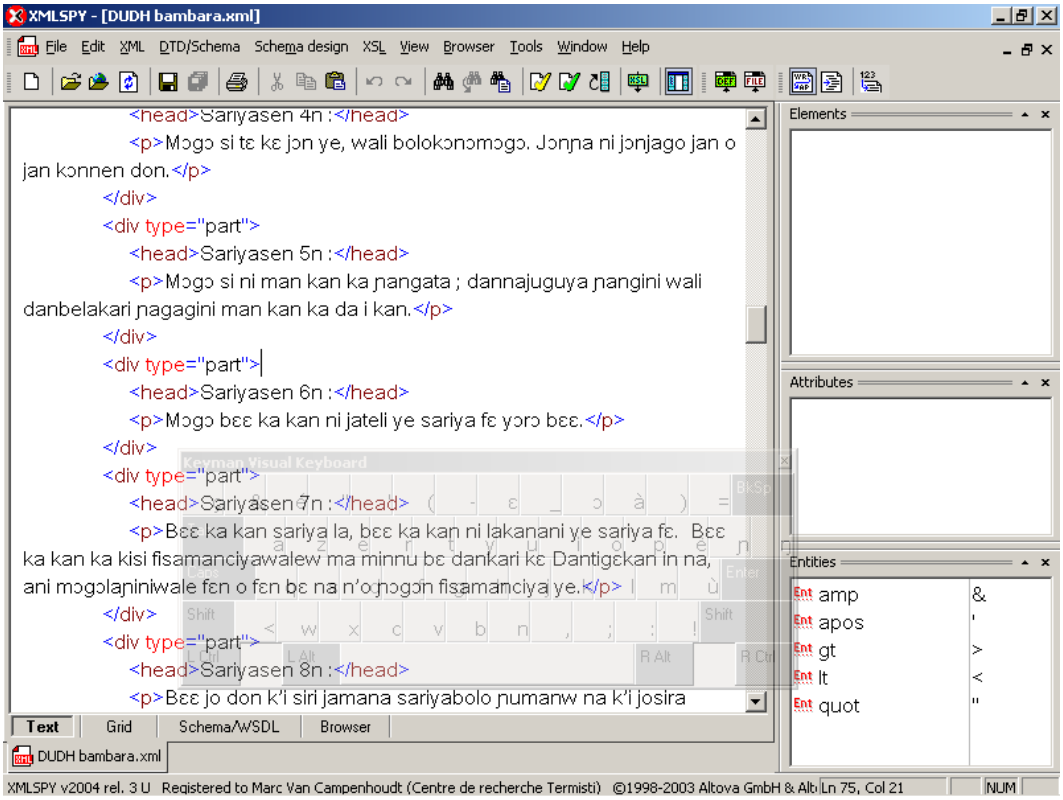

Figure 2 : encodage à l'aide du clavier virtuel bambara dans le logiciel *XML Spy*

11. www.tavultesoft.com/keyman.

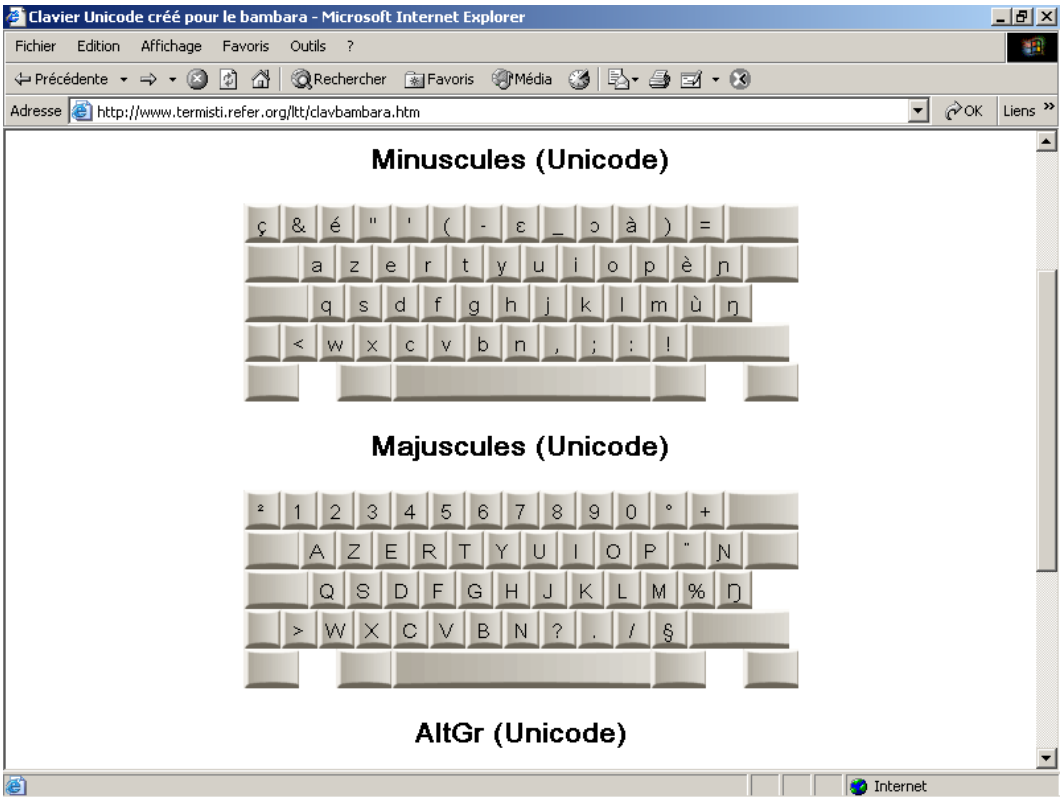

Figure 3 : documentation du clavier virtuel bambara générée en HTML

3.2.4 CONDITIONS D'UTILISATION DU CLAVIER VIRTUEL

Les premiers tests effectués montrent que le logiciel expérimenté fonctionne convenablement avec tous les logiciels réellement compatibles Unicode. Parmi ceux-ci, on citera, bien entendu, des suites bureautiques largement répandues comme *Office 2000* (on songe à *Word* et à *Access*), *Open Office* (gratuit) ou *Star Office* (gratuit pour l'enseignement). Plus spécifiquement, on citera aussi *XML Spy* ainsi que *Toolbox* (gratuit), la nouvelle version du célèbre progiciel *Shoebox,* utilisé de longue date pour la description des langues africaines.

Malheureusement, alors que la traduction constitue un enjeu majeur pour le développement des langues partenaires, les postes de travail du traducteur ne semblent pas encore tous parfaitement supporter Unicode ou, tout au moins, permettre l'encodage à l'aide d'un clavier virtuel. Il paraît donc quelque peu prématuré d'envisager un usage de ces produits pour des langues nécessitant l'emploi de caractères phonétiques. Par ailleurs, leur coût prohibitif demeure un obstacle important à leur usage en dehors de bureaux de traduction privés, d'organismes officiels ou d'ONG aux reins solides. Il existe, certes, l'un ou l'autre logiciel libre compatible Unicode, comme la mémoire de traduction *Omega T*[12], mais

12. www.omegat.org/omegat/omegat.html.

leur configuration et leur maniement demandent une excellente maîtrise de l'outil informatique.

3.3 Premières remarques concernant Unicode

Unicode est le fruit d'une collaboration internationale. Il est indéniable qu'à l'exception d'idiomes qui, comme l'amharique, se distinguent par un alphabet particulier, l'écriture des langues africaines n'a pas été réellement prise en compte par cette norme. Le consortium Unicode semble considérer que celles-ci ne possèdent pas une écriture propre et que l'usage combiné des caractères latins, de l'alphabet phonétique international et des diacritiques suffit amplement à leur représentation. Ceci pose parfois le problème de l'absence d'un caractère (p.ex., la capitale correspondant à la lettre minuscule latine « ʉ » barré) et complique surtout le rendu des glyphes[13] lors de la mise en page (p.ex. le « g » accompagné d'un diacritique rond souscrit (g̥ = g + ◌̥) est mal rendu dans les polices expérimentées[14]). L'indication correcte des tons, précieuse dans l'édition scientifique ou scolaire, est sensiblement compliquée en l'absence de table présentant tous les caractères imaginables en termes de combinatoire. Par ailleurs, Unicode ne permet pas, à cette date, une bonne représentation des tons modulés.

On est en droit de se demander pourquoi Unicode intègre des caractères accentués de langues européennes (p.ex. le « ê »), réputés équivalents à des combinaisons de lettres simples et de diacritiques (ê = e + ^), mais oblige le rédacteur des langues africaines considérées à recourir systématiquement à des combinaisons de caractères sans garantie de rendu du glyphe correspondant. Pourtant, Unicode contient parfois plusieurs fois le même caractère[15] et ne doit faire face à aucun besoin d'économie : des milliers de positions de code demeurent disponibles. En fait, on aborde ici une problématique récurrente en matière de normalisation internationale : la sous-représentation des pays du Sud au sein des comités techniques qui se chargent de l'élaboration de normes. Cette sous-représentation n'est pas tant due à un déni d'existence ou à l'absence de considération par les lois du marché – la variété des écritures rares incorporées dans Unicode en témoigne - qu'à l'absence de financement des missions de personnes compétentes lors des réunions de ces comités. Un travail de fond pourrait assurément être réalisé si les organisations internationales qui se soucient de l'adaptation des langues partenaires aux autoroutes de l'information considéraient le financement de telles missions comme une nécessité fondamentale. Il existe, certes, une procédure pour proposer l'inclusion de nouveaux caractères via Internet, mais elle suppose, outre une bonne connaissance de l'anglais, une excellente maîtrise de la problématique du codage des caractères et la volonté tenace de faire progresser un dossier.

---

13. On distingue normalement le caractère (abstrait) du glyphe, qui est sa représentation concrète dans telle ou telle police (Andries 2002 : 60-61).

14. Le nouveau format de fichier de polices multiplateforme *OpenType®* (Adobe et Microsoft) de même que le projet *Graphite* (Sil) devraient permettre une plus grande précision typographique.

15. P. ex., la *lettre majuscule latine A rond en chef* (U+00C5) correspond au symbole *angström* (U+212B) : Å (Andries 2002 : 69).

4 PERSPECTIVES

Au terme de quelque six mois de travail, les partenaires de l'action de recherche ont déjà pu faire la démonstration qu'un balisage XML des langues partenaires est parfaitement réalisable pour une variété de données textuelles et lexicales : il permet une fine représentation des contenus et une représentation universelle des caractères. L'exploration des possibilités offertes par l'écriture de feuilles de style – sans rentrer dans de la programmation – devrait ultérieurement permettre de vérifier si le langage XSL permet au linguiste de mieux valoriser son travail sur la toile.

Si l'on peut penser que le renouvellement, même lent, du parc informatique permettra à l'ensemble des chercheurs du Sud d'accéder à Unicode, une difficulté majeure continuera sans doute à freiner les efforts : la mauvaise connaissance de la langue anglaise – héritage malheureux de la colonisation - empêche nombre de linguistes de s'approprier des connaissances relativement aisées à maîtriser. Un effort majeur de la Francophonie devrait porter sur la traduction rapide en langue française des logiciels et des normes[16]. Le constat que toute l'information utile au balisage et au codage des langues n'est disponible qu'en anglais (voir la bibliographie), alors que les chercheurs francophones participent largement à leur création est tout à la fois révélateur des pratiques actuelles et lourd de conséquences pour la valorisation du patrimoine linguistique des pays du Sud.

Thierno Cisse,
Département de linguistique, Université Cheikh Anta Diop de Dakar.
Chérif Mbodj,
Centre de linguistique appliquée de Dakar, Université Cheikh Anta Diop de Dakar.
Marc Van Campenhoudt,
Centre de recherche Termisti, Institut supérieur de traducteurs et interprètes, Haute École de Bruxelles.
Mohamédoune (dit Doudou) Wane
Département des langues nationales et de linguistique, Université de Nouakchott.

16. Le centre de recherche Termisti a pris l'initiative de réaliser une traduction de la norme XCES, en accord avec ses concepteurs. Cette traduction devrait être rendue consultable dans les prochains mois.

BIBLIOGRAPHIE

Andries (P.), 2002 : « Introduction à Unicode et à l'ISO 10646 », dans *Document numérique*, 2002, vol. 6, n^{os} 3-4, p. 51- 88 (version mise à jour le 22 décembre 2003 : http://iquebec.ifrance.com/hapax/pdf/intro-Unicode.pdf)

Chanard (Chr.) et Popescu-Belis (A.), 2001 : « Encodage informatique multilingue : application au contexte du Niger », dans *Cahiers du Rifal,* décembre 2001, n° 22, p. 33-45.

Descotte (S.), Husson (J.-L.), Romary (L.), Van Campenhoudt (M.), Viscogliosi (N.), 2001 : « Specialized lexicography by means of a conceptual data base: establishing the format for a multilingual marine dictionary », dans Vainio (J.), éd., *Maritime Terminology : Dictionaries and Education, Proceedings of the Second Conference on Maritime Terminology, 11-12 May 2000, Turku,* Finland, Turku : University of Turku, p. 63-81 (Publications from the Centre for Maritime Studies, University of Turku, A36).

Dialo (A.), Mbodj (Ch.), Seck (A.N.) et Thiam (N.), 1997 : *Vocabulaire des élections (wolof-français suivi d'un index français-wolof), sous la direction de Chérif Mbodj, Dakar*, Centre de linguistique appliquée de Dakar.

Ide (N.) et Véronis (J.), 1996 : « Codage TEI des dictionnaires électroniques », dans *Cahiers GUTenberg, Numéro spécial : TEI - Text Encoding Initiative,* juin 1996, n° 24, p. 170-176.

Iso-CEI 8 859, 1998-2001 : *Information technology - 8-bit single-byte coded graphic character sets (Parts1-16)*, Genève : Organisation Internationale de normalisation. (disponible uniquement en anglais)

Iso-CEI 9 995-8, 1994 : *Information technology - Keyboard layouts for text and office systems - Part 8: Allocation of letters to the keys of a numeric keypad,* Genève : Organisation Internationale de normalisation. (disponible uniquement en anglais)

Iso-CEI 10 646, 2003 : *Information technology - Universal Multiple-Octet Coded Character Set (UCS),* Genève : Organisation Internationale de normalisation. (disponible uniquement en anglais)

Iso-CEI 14 755, 1997 : *Technologies de l'information - Méthodes de saisie de caractères du répertoire de l'Iso/CEI 10 646 à l'aide d'un clavier ou d'autres unités d'entrée*, Genève : Organisation Internationale de normalisation.

Iso 12 620, 1999 : *Aides informatiques en terminologie – Catégories de données*, Genève : Organisation internationale de normalisation.

Iso 16 642, 2003 : *Computer applications in terminology - Terminological markup framework,* Genève : Organisation internationale de normalisation. (disponible uniquement en anglais)

Mbodj (Ch.), 2002 : *Pour une terminologie de la santé en wolof*, Dakar : Centre de linguistique appliquée de Dakar.

Sperberg-McQueen (C.M.). et Burnard (L.), éd., 2002 : *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Oxford, Providence, Charlottesville, Bergen : Text Encoding Initiative Consortium. XML Version (www.tei-c.org).

Van Campenhoudt (M.), 2002 : « Lexicographie *vs* terminographie : quelques implications théoriques du projet DHYDRO », dans Zinglé (H.) dir., *Travaux du Lilla*, Université de Nice-Sophia Antipolis, 2002, n° 4, p. 91-103.

RÉSUMÉ

Cette communication présente les premiers enseignements de l'action de recherche en réseau *Expérimentation de normes de balisage en langues partenaires*. Les auteurs s'intéressent au balisage selon la norme XML de corpus lexicaux et textuels dans différentes langues du Sénégal et de Mauritanie (balante, bambara, pulaar, serer et wolof). Ils abordent également la problématique de l'écriture de ces langues à l'aide de claviers de saisie virtuels permettant d'utiliser le standard Unicode.

Mots clés : XML, XCES, Unicode, claviers virtuels, langues partenaires.