# Élaborer des corpus XML en langues partenaires : quelles technologies appropriées ? Une expérience en Mauritanie et au Sénégal

**Résumé**

L'élaboration de corpus textuels informatisés représente un défi majeur pour le développement des langues partenaires. Les technologies Unicode, XML et XSL ont, certes, des potentialités séduisantes, mais leur maîtrise effective peut se révéler délicate face aux réalités du terrain. S'inspirant des notions de technologies et méthodologies appropriées au développement durable, les partenaires de l'action de recherche *Expérimentation de normes de balisage en langues partenaires* proposent d'envisager une chaîne de traitement qui soit tout à la fois simple, efficace et aisément reproductible. Ils soulignent particulièrement l'enjeu primordial de la graphisation et l'obstacle que constitue la langue anglaise, *lingua franca* de l'ingénierie linguistique.

Mots clés : corpus, XML, graphisation, langues partenaires, technologies appropriées, Sénégal, Mauritanie.

## 1 Introduction

L'action de recherche partagée *Expérimentation de normes de balisage en langues partenaires*[1], financée par le réseau Lexicologie, terminologie, traduction (LTT) de l'Agence universitaire de la francophonie (AUF), a eu pour ambition de permettre le balisage XML en écriture Unicode de bases de données textuelles et lexicales. L'idée était « de vérifier l'intérêt des propositions de normes existantes au regard de la réalité de langues souvent négligées par la normalisation internationale » (Cisse *et al.* 2004 : 82) en vue de diffuser sur les inforoutes des données linguistiques propres aux langues du Sénégal et de Mauritanie.

La présente communication revient sur les difficultés rencontrées au cours de cette recherche appliquée, alors que nous ne mesurions pas toutes les réalités du terrain. Le poids et l'influence des travaux des universitaires du Nord étaient bien connus, l'un de nos objectifs étant précisément de montrer qu'ils ne constituaient pas une fatalité. Par contre, nous avions moins bien mesuré la difficulté de récupérer et réunir des corpus textuels ou lexicaux déjà informatisés. Plus encore, les contraintes liées aux différentes fractures qui frappent les pays du Sud se sont révélées très prégnantes, au point de nous pousser, au terme de la recherche, à reconsidérer notre approche au travers du prisme des technologies et méthodologies appropriées au développement durable.

1. www.termisti.refer.org/ltt/ltt.htm. Partenaires : Université Cheikh Anta Diop de Dakar (Centre de linguistique appliquée de Dakar et département de linguistique), Université de Nouakchott (Département des langues nationales et de linguistique) et Institut supérieur de traducteurs et interprètes (Haute École de Bruxelles, Centre de recherche Termisti).

## 2 Les technologies et méthodologies appropriées au développement durable

### *2.1 Les technologies appropriées*

La notion de technologies appropriées au développement a fait florès dans le monde des ONG au tournant des années 70-80. Elle est née de l'échec des grands programmes de coopération mis en place dans le sillage de la décolonisation. Le mouvement des technologies appropriées, qui s'est étendu aux questions de méthodologie, s'appuie largement sur l'idée d'un microdéveloppement fondé sur les moyens immédiatement disponibles : « Toute société dispose de technologies qui assurent son développement ou du moins sa survie. Ces technologies sont le résultat de la capacité d'invention et d'adaptation de cette société. Leur degré de sophistication et de complexité varie considérablement d'une société à une autre pour des raisons multiples. » (Crombrugghe 1984 : 65.)

Mieux qu'une brève définition, une série non exhaustive de critères permettent d'évaluer l'appropriation d'une technologie aux besoins en développement durable : « l'économie de devises, l'économie d'investissement, l'intensité en main-d'œuvre, l'économie d'énergie, l'utilisation d'énergies renouvelables, la décentralisation de la création de valeur ajoutée, la préservation du potentiel écologique, l'autonomie technique et financière des utilisateurs pour le fonctionnement, l'entretien et la réparation des équipements, la reproductibilité locale, le potentiel de diffusion, l'acceptation par les populations, l'utilisation des matériaux locaux, l'utilisation du savoir-faire et de l'expérience locale. » (Crombrugghe 1984 : 65[2].)

Le mouvement s'est surtout focalisé sur des besoins fondamentaux du développement : l'eau, la terre, l'énergie, la santé. À ce jour, très peu de travaux semblent avoir porté sur la graphisation, la bureautique ou encore la microédition.

### *2.2 Le développement durable*

Soucieuses de l'homme et respectueuses de l'environnement, les techniques et méthodologies appropriées se sont inscrites, avant l'heure, dans la perspective du « développement durable ». Cette dernière notion a été mise en avant en 1987 par les Nations unies au départ du « rapport Brundtland » (Cmed 1988). En 1992, la Conférence des Nations unies sur l'environnement et le développement (Cnued) a adopté à Rio vingt-sept principes devant permettre un tel développement et établi un guide de mise en œuvre du développement durable au 21e siècle, le fameux *Agenda 21*.

2. Voir aussi les critères énoncés par Darrow et Saxenian (1993).

### 2.2.1 Le rôle de la langue commune et des langues partenaires

Le rôle primordial des langues partenaires n'a guère été souligné dans le cadre de cette dynamique et la question linguistique n'est guère abordée de manière spécifique dans les grands textes de référence[3]. La Francophonie a toutefois le mérite particulier d'avoir intégré la question de la diversité culturelle et linguistique lors du récent colloque *Développement durable : leçons et perspectives* (Ouagadougou, 1er-4 juin 2004). On est particulièrement heureux de lire dans le résumé des principales recommandations que « les participants au colloque réaffirment le caractère inaliénable de la diversité culturelle et linguistique comme fondement du développement durable »[4].

### 2.2.2 Développement durable et fracture numérique

Les débats concernant les différentes fractures - médicale, éducative, sociale, numérique… - qui affectent le développement posent, certes, la question de l'adéquation des TIC aux besoins réels (*cf.* Guichard 2003 et Valérien et Wallet 2004 : 119). On conviendra toutefois que l'ingénierie linguistique offre aujourd'hui des possibilités extraordinaires de description des langues partenaires, souvent en danger, et permet d'envisager une communication efficace entre locuteurs de ces langues. Le lien entre le développement des langues partenaires et celui des technologies de l'information apparaîtra évident à la communauté des linguistes, alors qu'il n'a pas toujours été explicité, loin s'en faut, dans les grandes déclarations officielles[5].

## *2.3 Ingénierie linguistique et technologies appropriées*

Notre expérience nous amène à souligner combien les conditions de travail des chercheurs du Sud compliquent l'émergence d'une ingénierie linguistique autonome. Il nous semble donc primordial de définir les technologies et méthodologies les plus adéquates pour leur permettre de contribuer au mieux à l'étude et au développement de leurs langues, conçues comme les médiums appropriés d'un développement durable. Nous n'entendons pas évoquer ici la nécessité d'un matériel informatique approprié, exploitant des logiciels libres et dont l'énergie serait fournie par une manivelle[6]. Il s'agit plus modestement d'envisager la méthodologie logicielle la plus appropriée à notre objectif initial de création de corpus linguistiques au Sénégal et en Mauritanie. Notre réflexion nous conduit ainsi à envisager divers critères d'évaluation des voies les plus appropriées à cet objectif. Au regard de celui-ci et des situations locales, tous ces critères n'auront bien sûr pas le même poids mais peuvent constituer autant de repères au moment d'opter pour la mise en œuvre d'une chaîne de

---

3. Le *Glossaire pour le développement durable,* qui propose une traduction en français des termes les plus utilisés ne contient quasiment aucun terme lié aux sciences du langage (www.francophonie-durable.org/textes.html).

4. www.francophonie-durable.org/documents/Principales_recommandations.pdf.

5. Voir, par exemple, le plan d'action issu de la *3e Conférence ministérielle sur la culture* organisée par l'AIF à Cotonou les 14 et 15 juin 2001.

6. Pour ce type de problématique, voir le *Hundred-Dollar Laptop Project* Media Lab (MIT) : laptop.media.mit.edu.

traitement de corpus[7]. Certains sont d'ordres technique et financier : gratuité ou faible coût des logiciels ; installation depuis un support « ancien » (disquette, cédérom) ; Internet non indispensable ; fonctionnement multiplateforme, même sur un système daté ; faible consommation de ressources système (mémoire vive, espace disque, etc.). D'autres concernent les fonctionnalités mêmes des logiciels : compatibilité avec la norme Unicode ; entrées et sorties en formats non propriétaires et, dans la mesure du possible, stockage interne dans un format ouvert (p.ex. XML) ; modularité et intégration des fonctionnalités (le même logiciel traite plusieurs étapes).

L'aspect humain doit également être considéré. Il importe de garantir l'autonomie d'un chercheur qui ne peut ni recourir à un informaticien ni s'improviser informaticien. Il convient donc d'éviter, d'une part, qu'il doive effectuer de nombreuses manipulations complexes de données au travers d'une succession de logiciels distincts et, d'autre part, que la prise en main de chaque outil se révèle ardue. À cet égard, un critère essentiel de l'appropriation est la disponibilité de l'interface et de la documentation (hors ligne) des outils dans la langue commune (le français)[8].

## 3 Accéder à l'information en langue française

Les ONG actives dans le domaine des technologies appropriées se sont toujours souciées de la communication des savoirs pratiques, au travers de centres de documentation ou sous la forme de fiches simplifiées, rédigées dans un français accessible. Certaines, comme l'Inades (Institut africain pour le développement économique et social), ont particulièrement été attentives à l'utilisation du français fondamental. Des manuels ont parfois même été rédigés en langues partenaires (p.ex. Schmitz 1986).

### *3.1 Une* lingua franca *incontournable*

L'anglais est la langue véhiculaire de l'informatique. La très grande majorité des documents indispensables pour s'approprier les technologies de l'ingénierie linguistique ne sont disponibles qu'en anglais[9], alors même qu'ils ont parfois été rédigés par des francophones. La seule manière de participer au débat scientifique – par exemple pour faire évoluer la norme Unicode (*cf.* 4) – est de s'exprimer dans cette langue. Force est, malheureusement, de constater que celle-ci demeure très mal maîtrisée par nombre de diplômés du Sud éduqués dans l'idée de la grandeur de la langue métropolitaine. Ne pouvant lire des documents indispensables, ils se trouvent très rapidement bloqués dans leurs possibilités d'élaborer des méthodologies fondées sur le génie logiciel. Un rapide inventaire des normes et programmes disponibles prouve vite qu'il est

7. Nous assumons le fait que certains de ces critères sont incompatibles : l'émergence d'Unicode est liée au progrès informatique et suppose un système d'exploitation récent. Des passerelles avec des systèmes plus anciens sont heureusement possibles (Chanard et Popescu 2001).

8. Ce critère est une condition nécessaire mais non suffisante d'appropriation : quand bien même une interface et une documentation sont disponibles en langue commune, il convient ensuite d'en vérifier la qualité sous l'angle de la forme (qualité linguistique) et sous celui du fond (richesse documentaire, p.ex. disponibilité d'un didacticiel, de fichiers d'exemples, d'un manuel de l'utilisateur et d'un manuel de référence, etc.).

9. Norme Unicode, normes du W3C, normes d'échange de données linguistiques, logiciels de balisage, concordanciers…

impossible de travailler sans bien comprendre l'anglais. Des traductions fragmentaires existent parfois, mais elles sont le plus souvent le fruit d'initiatives dispersées et le fait de bénévoles non formés à la traduction. En outre, les mises à jour des documents originaux sont rarement prises en compte.

### *3.2 Pour une cellule de traduction « inforoutes »*

Il serait vain, sinon puéril, de se battre pour obtenir que le français devienne la langue internationale du progrès informatique. Il est, bien sûr, attristant de constater que des chercheurs francophones ne diffusent pas également dans leur propre langue les normes internationales qu'ils ont rédigées, nécessité fait loi, en anglais. Il serait toutefois injuste de leur faire le procès de ne pas se soucier de l'existence d'une version en langue française : la traduction exige, outre une formation particulière, de disposer de moyens logistiques et financiers adéquats. Imposer le français sur le théâtre de la conception des autoroutes de l'information est un combat déjà perdu. Assurer une traduction professionnelle rapide permettrait, par contre, d'éviter qu'un fossé encore plus important ne se creuse entre le Sud et le Nord.

Les linguistes francophones n'occuperaient-ils pas davantage le terrain de la normalisation des échanges de données s'ils pouvaient disposer très rapidement d'une traduction officielle de toutes les publications importantes ? La Francophonie institutionnelle serait assurément bien inspirée de financer un programme de traduction de l'anglais de tous les documents nécessaires à une appropriation rapide des savoirs indispensables à la création des inforoutes en français et en langues partenaires[10]. Traduire ensuite vers l'anglais les réactions des linguistes du Sud permettrait d'interagir rapidement avec les créateurs de normes pour veiller à la prise en compte des besoins propres aux langues partenaires. Par ailleurs, la localisation des quelques logiciels incontournables pour la description linguistique faciliterait grandement le travail de nombreux chercheurs qui n'arrivent guère à en exploiter toutes les potentialités[11]. Que coûterait le financement annuel d'une cellule légère de deux ou trois traducteurs délocalisée au Sud au regard du coût de l'organisation d'événements qui tiennent plus de la vitrine éphémère que du développement durable ?

## 4 La problématique de la graphisation

La graphisation des langues partenaires constitue un enjeu majeur. Elle est la condition *sine qua non* de leur traitement informatique et donc de leur utilisation sur le réseau Internet. Certains proposent de simplifier leur écriture pour éviter les difficultés de numérisation générées par les diacritiques et autres signes moins fréquents (voir, par exemple, Chaudenson 2004). On devine aisément que les caractères empruntés à l'alphabet phonétique international sont les premiers visés par les tenants de ce point de vue.

---

10. Une retombée indirecte mais non négligeable de l'activité d'une telle cellule pourrait être la standardisation de la terminologie de ce sous-domaine des systèmes d'information, en coopération avec les organismes francophones compétents.

11. Nous pensons particulièrement au célèbre logiciel *Toolbox* (ex-*Shoebox*).

Cette proposition semble, *a priori,* inspirée par la philosophie des technologies appropriées. Elle néglige malheureusement la question fondamentale de l'appropriation : on ne peut nier aux locuteurs des langues partenaires le droit de choisir, eux aussi, la meilleure manière d'écrire leur langue. Les peuples européens ont-il abandonné leurs alphabets pour se conformer à la norme américaine ASCII fondée sur les 26 lettres de l'alphabet latin ? La loi du marché et les progrès technologiques ont plutôt débouché sur une adaptation des méthodes de codage aux besoins des consommateurs : la norme Unicode, censée inclure toutes les écritures du monde dans une table unique, sans nécessité de procéder à des transcodages.

### *4.1 Les réalités du terrain*

Au Sénégal et dans les pays de la même aire linguistique, il existe une longue tradition de description des traditions orales et de constitution de corpus oraux. Les textes législatifs sur l'écriture des langues nationales sénégalaises datent des années 70 et depuis, il n'y a jamais eu de politique ardue de développement des langues locales et d'incitation à l'écriture standardisée. Le français demeure la langue officielle et c'est seulement depuis une décennie, avec le développement des radios libres, que les langues locales, et notamment le wolof, ont commencé à concurrencer le français, mais uniquement au niveau discursif.

Les premières transcriptions n'ont pas toujours tenu compte des besoins de constitution de corpus écrits d'envergure et de partage des données ; les alphabets utilisés sont différents d'un chercheur à l'autre. Une graphisation harmonisée à l'échelle des aires linguistiques qui permette une communication scientifique et une présence sur Internet demeurent indispensables pour la survie des langues et cultures locales. Il s'agit de reprendre systématiquement les transcriptions existantes et en même temps de vulgariser les outils modernes de prise en charge des langues locales.

La solution de la graphisation harmonisée, si elle veut être opérante, doit tenir compte des problèmes logistiques : l'électricité déficiente, l'équipement informatique réduit, souvent vieillissant et fort différencié d'un laboratoire à l'autre dans le même pays et d'un pays à l'autre, les antivirus inefficaces, les connexions lentes – lorsqu'elles existent - et la pénurie de personnel informatique affecté aux recherches en sciences humaines et sociales. Cette solution tiendra ainsi compte de la fameuse « fracture numérique » dans le choix des logiciels et des technologies les plus appropriées, respectueuses de l'expression locale.

### *4.2 Résoudre la question du clavier, une étape indispensable*

L'absence de véritable marché explique vraisemblablement la non-commercialisation de claviers d'ordinateur propres aux grandes langues africaines. La faiblesse économique explique sans doute également que les écritures des langues d'Afrique de l'Ouest ne sont pas encore intégrées comme telles dans Unicode. L'indigence pécuniaire n'est cependant pas nécessairement un obstacle : on devine que seul un intense travail de *lobbying* a permis l'intégration dans Unicode des hiéroglyphes, des syllabaires autochtones canadiens, du syllabaire éthiopien et, depuis peu, de l'écriture tifinagh du berbère. Une récente démarche, soutenue par le monde anglophone, est en passe de faire aboutir l'intégration de l'écriture n'ko du mandingue dans Unicode, preuve de la nécessité de soutenir des initiatives volontaristes. La francophonie pourrait aisément financer la participation régulière d'experts africains aux activités du consortium Unicode. Ceci supposerait, encore une fois, que soit créée une cellule de traduction apte à soutenir les chercheurs travaillant sur la graphisation des

langues partenaires (*cf.* 3.2), à moins de préférer qu'ils ne s'associent aux universités d'outre-Atlantique, très attentives à la question[12].

### *4.3 Le clavier virtuel, solution simple et bon marché*

La possibilité de réaffecter les touches d'un clavier physique est une procédure alléchante. Dans le cadre de notre projet, deux logiciels ont été expérimentés : *Keyman Developper* et *Microsoft Keyboard Layout Creator (MKLC)*[13]. Comme les partenaires du projet l'ont déjà montré (Cisse *et al.* 2004 : 86-92), les claviers créés[14] fonctionnent dans un grand nombre d'applications courantes, offrent la certitude d'utiliser le bon caractère Unicode, sont gratuits, tiennent sur une disquette et s'installent aisément. Il n'est pas nécessaire de connaître Unicode pour les utiliser au quotidien, mais leurs interfaces anglaises devraient être localisées en français.

La création de ce type de clavier suppose d'identifier les caractères adéquats dans Unicode, en l'absence d'un véritable « bloc » qui soit consacré aux langues concernées. Le petit gratuiciel *Babelmap,* dont l'archive pèse à peine 600 k et qui ne nécessite aucune connexion à Internet, constitue sans doute une solution simple et maniable pour qui souhaite s'atteler à cette tâche. Disponible en français, il propose également une documentation Unicode dans cette langue[15].

Le processus de création de claviers proprement dit ne devrait intéresser *a priori* que des chercheurs en linguistique. Le gratuiciel *MKLC* s'est révélé plus simple que *Keyman Developper,* même si les potentialités de ce dernier sont très intéressantes pour créer un clavier très étendu. Malheureusement, ces deux interfaces et leurs documentations n'existent qu'en anglais.

Les claviers créés à l'aide de *Keyman Developper* ou de *MKLC* sont relativement simples d'emploi, mais les premiers se révèlent parfois capricieux. Les claviers *MKLC* sont intégrés à *Windows* : ils peuvent donc s'installer dans une interface française et s'utiliser selon la classique procédure de basculement de clavier propre à ce système d'exploitation. Il importe toutefois d'observer que si *Keyman* et *MKLC* fonctionnent avec les programmes de bureautique les plus courants, ainsi qu'avec des éditeurs XML comme *XML Spy* et *Oxygen,* certains logiciels bien connus ne les acceptent pas nécessairement[16].

---

12. Voir, par exemple, la très intéressante *Script Encoding Initiative* de l'Université de Californie (Berkeley) : www.linguistics.berkeley.edu/sei/index.html.

13. www.tavultesoft.com/keymandev et www.microsoft.com/globaldev/tools/msklc.mspx.

14. Téléchargeables à l'adresse : www.termisti.refer.org/ltt/ltt03.htm

15. www.babelstone.co.uk/Software/BabelMap.html.

16. Tel est le cas du célèbre éditeur HTML *Web Expert 6.* Les claviers *MKLC* ne semblent pas fonctionner dans le logiciel *Toolbox*, lequel fonctionne très bien avec *Keyman*.

## 5 Quelle chaîne de traitement du corpus textuel ?

Notre objectif fondamental était la mise en commun de données linguistiques en langues partenaires grâce à l'usage conjoint d'Unicode et d'une norme XML. Nous pensions que le mécanisme de transformation des documents XML à l'aide du langage de feuilles de style XSL permettrait, dans un second temps, une diffusion aisée des textes sur la toile. Dans ce cadre, les standards d'échange XCES (*Corpus Encoding Standard for XML*) et TEI (*Text Encoding Initiative*), qui permettent d'enrichir le document par une grande variété de descripteurs, ont été principalement testés.

### *5.1 L'utilisabilité des normes XML*

L'échange de données lexicales selon le standard XML semble relativement aisé à maîtriser dès lors que le modèle de données a été bien pensé et que les catégories de données sont restreintes[17]. La maîtrise des normes textuelles XCES et TEI s'avère sans doute plus délicate au vu de la masse d'informations à appréhender. La concurrence entre les normes TEI et XCES est regrettable dans la mesure où leurs différences ne semblent pas évidentes au premier abord. L'impression de solitude est grande pour qui souhaite s'initier en autodidacte à ces normes et arriver à une bonne maîtrise. Quel que soit le contenu à baliser, le principal écueil réside, encore une fois, dans la nécessité de bien comprendre la langue anglaise. Seule la TEI *Lite* possède une traduction officielle, particulièrement bienvenue[18] : le souhait d'aller à l'essentiel en fait un bon outil en termes de technologies appropriées, même si l'on doit regretter que la simplification ait réduit la portée de la norme aux textes « littéraires », sans véritable prise en compte des catégories propres à la littérature orale. Par ailleurs, la volonté des concepteurs de la TEI de proposer une interface capable de générer une « sous-DTD » adaptée à des besoins spécifiques[19] est particulièrement louable.

### *5.2 Éditeurs XML spécifiques*

On trouve aisément sur la toile des éditeurs XML gratuits ou à coût réduit. Outre que beaucoup n'offrent que peu de fonctionnalités et souffrent de l'absence d'une documentation exhaustive, la plupart ne sont, ici encore, disponibles qu'en anglais. L'action de recherche ne visant pas à les expérimenter spécifiquement, les partenaires ont travaillé successivement avec deux produits renommés et de coût abordable : *XML Spy* et *Oxygen*[20]. Le deuxième présente le double avantage d'être localisé et documenté en français et de proposer des feuilles de style adaptées au traitement de documents XML conforme à la TEI.

---

17. Le traitement des données lexicales et terminologiques, également abordé dans le cadre de la recherche, n'est pas envisagé ici.

18. *La TEI simplifiée* : www.tei-c.org/*Lite*/teiu5_fr.html.

19. www.tei-c.org/pizza.html.

20. Hormis les deux éditeurs commerciaux susmentionnés, divers logiciels à code ouvert (*Bitflux, Jaxe, jEdit, Peter's XML Editor, TreeBeard* et *Vex*) ont été mis à l'essai mais ont été écartés car ils ne satisfaisaient pas à un ou à plusieurs des critères d'appropriation évoqués.

Le public-cible de ces logiciels est relativement restreint. Les linguistes qui ne sont pas informaticiens mais connaissent XML n'y auront vraisemblablement recours que pour disposer de facilités d'encodage, de validation du document et de conversion à l'aide de feuilles de style. Les points essentiels pour le choix d'une interface seraient donc logiquement : être disponible en français, répondre aux critères de choix des logiciels en termes de technologies appropriées, faciliter le travail de balisage, permettre d'associer aisément une DTD et une feuille de style, posséder un moteur de transformation garantissant le rendu des caractères Unicode propres aux langues traitées[21].

Par ailleurs, le langage de transformation XSL s'avère rapidement complexe, ce qui constitue un obstacle majeur au regard du critère d'autonomie des chercheurs du Sud. Il importe donc que soient rendues disponibles davantage de feuilles de style permettant de valoriser les corpus balisés. Malheureusement, les concepteurs des normes XCES et TEI ne semblent guère concernés par cette tâche, qui échappe à leur activité de normalisation.

### *5.3 Traitement de texte proposant des portes d'entrée et de sortie vers XML*

L'expérience nous donne à penser qu'un éditeur XML n'est pas un outil suffisamment simple pour nombre de linguistes qui engrangent des textes sans souhaiter nécessairement exploiter un balisage XML fin. Plutôt que de leur demander un balisage lourd et fastidieux, il semble préférable de leur proposer d'encoder leur corpus dans un traitement de texte selon une logique de document structuré : utiliser des styles renvoyant aux catégories de données *ad hoc* d'une norme d'échanges devrait permettre ensuite une conversion vers un format XML. L'expérience nous suggère que cette solution satisfairait nombre de linguistes ou d'étudiants en sciences du langage qui ne peuvent s'investir dans des technologies de l'information trop complexes.

Le consortium de la *Text Encoding Initiative* propose déjà une semblable procédure pour le traitement de texte *Writer* d'*OpenOffice.* Elle permet de sauvegarder un fichier *Writer* au format XML de la TEI et, inversement, d'ouvrir un fichier XML dans *Writer.* L'expérimentation montre la viabilité de cette solution qui présente le triple avantage d'être gratuite, aisée et peu gourmande en ressources informatiques. Certes, les fichiers proposés (DTD et XSL) ne dépassent pas la simple démonstration. Cependant, il semble tout à fait possible pour un spécialiste de les améliorer en veillant à diversifier les éléments présents dans la DTD et les styles qui leur correspondent dans le modèle de document proposé[22].

La plateforme *Cyberdocs* du projet *Cyberthèses*[23]*,* qui publie sur la toile des thèses du monde entier, exploite une approche similaire. Développée avec l'aide de l'Agence de la Francophonie, elle permet, en effet, de convertir vers la TEI *Lite,* et via *OpenOffice*, des fichiers de traitements de texte. Nous n'avons malheureusement pas pu vérifier si une telle solution garantit bel et bien un rendu correct des caractères Unicode (Abdrahamane 2004). À l'instar de *Cyberthèses*, l'Agence bibliographique de l'enseignement supérieur a mis au point un système de publication de thèses en ligne basé sur la TEI (*Sparte*[24]) ; à la différence toutefois que le

21. Le moteur de transformation propre à la version 2004 de *XML Spy Home Edition* s'est avéré décevant de ce dernier point de vue.

22. Cette remarque concerne particulièrement les passages versifiés et dialogués.

23. sourcesup.cru.fr/cybertheses.

24. www.abes.fr/abes/DesktopDefault.aspx?tabid=315.

format supporté avant la transformation XML est le format RTF. Divers autres projets francophones ne mettent pas en œuvre un balisage TEI avec *OpenOffice*, mais fournissent des modèles d'implémentation du couple *OpenOffice / XML* reposant sur les principes exposés ici et démontrent la faisabilité de la démarche proposée. Citons à titre de référence *Ooo2Dbk,* une chaîne de production documentaire mise en œuvre au sein du ministère français de l'Équipement[25] ou la plateforme de publication en ligne *Lodel*[26].

## 6 En guise de synthèse : quelques mesures pratiques

### *6.1 La mise en ligne de fiches pratiques*

Tout linguiste confronté à des problèmes informatiques a un jour éprouvé la difficulté de trouver une information cohérente, simple, pratique et rédigée en français sans devoir parcourir un grand nombre de sites aux réponses fragmentaires. Le mouvement des technologies appropriées a souvent veillé à transmettre ses connaissances à travers des fiches pratiques (on songe aux célèbres fiches du Gret[27]) qui ont parfois été informatisées (p.ex. Agridoc[28]). En ingénierie linguistique, de telles fiches, rédigées en un français accessible, permettraient à nombre de chercheurs, d'enseignants et d'étudiants du Sud comme du Nord de se débrouiller au quotidien pour implémenter des logiciels simples, disposer des polices Unicode adéquates, configurer correctement leur navigateur ou leur logiciel de courrier, produire un document XML, opérer une transformation XSL, etc.

Il serait également utile d'associer à de telles fiches pratiques des exemples concrets d'application aisément adaptables sur le terrain. On songera, par exemple, à des descriptifs Unicode des principaux caractères utilisés par les africanistes, à des claviers virtuels ou encore à des modèles d'application XSLT.

### *6.2 Des ressources francophones plus nombreuses et aisément disponibles*

La barrière de la langue anglaise plaide pour la mise en place d'un répertoire commenté des ressources francophones déjà disponibles en matière d'ingénierie linguistique. Une série, certes restreinte, de ressources francophones de qualité existent bel et bien sur la toile. Toutefois, leur visibilité est extrêmement faible parce qu'elles sont dispersées, résultent d'initiatives personnelles ou ne traitent qu'incidemment de contenus spécifiques à l'ingénierie linguistique. D'un coût anecdotique, une telle initiative permettrait de centraliser les informations indispensables à la valorisation des langues partenaires.

---

25. www.indesko.com/fr/telechargements/ooo2dbk.

26. www.lodel.org.

27. Groupe de recherche et d'échange technologiques (www.gret.org).

28. www.agridoc.com/fichestechniques_gret/index.htm.

Une politique volontariste de localisation des logiciels incontournables et de traduction des normes fondamentales disponibles uniquement en anglais – et de leurs mises à jour - permettrait de rapidement multiplier de semblables ressources. Son financement suppose un changement de stratégie dans la défense de notre langue commune : passer de la défense du pré carré francophone à un dialogue constructif et intelligent avec les locuteurs de la nouvelle *lingua franca* de la connaissance.

### *6.3 La diffusion des technologies Unicode*

Le clavier d'ordinateur constitue le point d'entrée du contenu numérique. Il ne coûterait guère de veiller à ce que les locuteurs des langues partenaires puissent écrire aisément leur langue à l'aide de l'outil informatique. Sans devoir miser sur des programmes prestigieux, il suffirait de soutenir des projets de diffusion de ces petits claviers virtuels Unicode auprès des milliers d'utilisateurs potentiels : écoles, centres de santé, municipalités, ministères, presse, gestionnaires de site Internet, ONG… Ils sont, en effet, une condition indispensable à la graphisation des langues, donc à leur utilisation sur un support informatique et sur les inforoutes. Aider à configurer en Unicode les logiciels les plus courants serait une mesure de soutien également aisée à mettre en œuvre.

### *6.4 La promotion du document structuré*

La maîtrise de XML n'est pas aisée pour qui souhaite créer des corpus et elle suppose une pratique régulière. Trop peu d'outils simples sont mis à la disposition des candidats utilisateurs et autres néophytes : les promoteurs des échanges structurés semblent préférer s'arrêter après une démonstration de faisabilité. Une utilisation quotidienne de ces technologies n'est donc viable que pour des institutions possédant une équipe d'informaticiens aptes à développer des interfaces simplifiées.

À défaut de pouvoir convertir la communauté des linguistes aux logiciels de la galaxie *LaTex,* qui produisent des documents structurés, il paraît plus sage de proposer des formations à une utilisation rationnelle des modèles de document de logiciels comme *Word* ou *Writer.* En effet, l'association d'un style à une catégorie de données permettant d'envisager une conversion vers un document XML (*cf.* 5.3), il semble plus réaliste de mieux segmenter les compétences : la tâche du linguiste s'arrêterait à l'engrangement selon un modèle de document conforme à une norme d'échange, celle du spécialiste XML étant de faciliter la conversion vers des standards internationaux et de créer des feuilles de style XSL immédiatement utilisables.

*Thierno Cisse*
*Département de linguistique*
*Université Cheikh Anta Diop de Dakar*

*Paul Muraille – Marc Van Campenhoudt*
*Centre de recherche Termisti*
*Institut supérieur de traducteurs et interprètes*
*Haute École de Bruxelles*

**Bibliographie**

Abdrahamane (A.), 2004 : « Cyberthèses : une solution à la visibilité de la science africaine ? », dans *Actes de la Conférence sur la publication et la diffusion électronique, Dakar, Sénégal, 1 - 2 septembre 2004,* Dakar : Codesria, www.codesria.org/Links/conferences/el_publ/Abdrahamane_Anne.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.

Chaudenson (R.), 2004 : « La graphisation des langues africaines », dans *Cahiers du Rifal*, décembre 2005, n° 24, p. 66-67.

Cisse (Th.), Mbodj (Ch.), Van Campenhoudt (M.) et Wane (M.), 2004 : « Expérimentation de normes de balisage en langues partenaires », dans *Actes des Premières Journées scientifiques communes des réseaux de chercheurs concernant la langue « Penser la Francophonie, concepts, actions et outils linguistiques »*, Université de Ouagadougou, 31 mai - 1er juin 2004, p. 81-93, consultable à l'adresse www.bibliotheque.refer.org/livre244/l24410.pdf.

Commission mondiale sur l'environnement et le développement (Cmed), 1988 : *Notre avenir à tous,* Montréal : Éditions du Fleuve - Les publications du Québec.

Crombrugghe (G. de), 1984 : « Favoriser le développement et l'appropriation de la technologie », dans *Le Courrier*, janvier-février 1984, n° 83, p. 65-66.

Darrow (K.) et Saxenian (M.), 1993 : *Appropriate Technology Sourcebook*, Standford : Volunteers in Asia. Consulté à l'adresse : villageearth.org/atnetwork/atsourcebook.

Guichard (É.), 2003 : « Does the 'Digital Divide' Exist? », dans Seters (P. van), Gaay Fortman (B. de) et Ruijter (A.), dir., *Globalization and its new divides: malcontents, recipes, and reform*, Amsterdam : Dutch University Press. Traduction française : barthes.ens.fr/atelier/geo/Tilburg.html.

Schmitz (J.-L.), 1986 : *Mutwisi ti bangombe na yandi. Bisalu ya kutwila bangombe na babwala na ndambu ya westi ya Zaïre,* Kinshasa : Inades-Formation, 221 p. (publié en français sous le titre *L'éleveur et son bétail. L'élevage bovin villageois dans l'ouest du Zaïre).*

Valérien (J.) et Wallet (J.), 2004 : « À quelles conditions un projet intégrant les TIC dans l'éducation peut-il être considéré comme au service du développement durable ? », dans *Actes du colloque Développement durable : leçons et perspectives (Ouagadougou 1er-4 juin 2004)*, p. 117-122. Consulté à l'adresse www.francophonie-durable.org/sommaire.html.

---