Apports d'Unicode à l'édition numérique multilingue

Une étude de cas au Niger

Andrei Popescu-Belis

Université de Genève, Ecole de Traduction et d'Interprétation
Institut pour les Etudes Sémantiques et Cognitives (ISSCO)
40, bd. du Pont d'Arve
CH-1211 Genève 4 – Suisse
andrei.popescu-belis@issco.unige.ch

RÉSUMÉ. *Cet article met en évidence l'apport d'Unicode au traitement des documents multilingues en Afrique, à partir d'une expérience concrète au Niger. Nous analysons d'abord les liens entre l'informatisation des langues et leur écriture, puis nous donnons un aperçu des difficultés que pose le multilinguisme à l'édition assistée par ordinateur, en étudiant le processus éditorial dans un institut de documentation nigérien. Nous montrons alors en quoi le standard Unicode peut rendre ce processus plus efficace, avant de décrire plusieurs solutions aux difficultés rencontrées, en fonction des outils informatiques disponibles actuellement et à l'avenir.*

ABSTRACT. *This article describes how Unicode can improve multilingual document processing, based on a concrete study conducted in Niger. We first analyze the links between the existence of a writing system and the computational processing of a language. Next, we outline the challenges of computerized multilingual editing, by studying the editing process in a Nigerien documentation institute. We show then how the Unicode standard can render this process more efficient, and describe several solutions to the observed difficulties, depending on the software and hardware available now and in the future.*

MOTS-CLÉS : *édition numérique multilingue, Unicode, langues nigériennes, polices de caractères, codage.*

KEYWORDS: *computerized multilingual editing, Unicode, Nigerien languages, fonts, encoding.*

Document numérique. Volume 6 – n° 3-4/2002, pages 139 à 153

1. Introduction

Le standard Unicode ou ISO/CEI 10646 (ci-après, « Unicode ») permet de faciliter la représentation informatique d'alphabets variés, et en particulier la rédaction et le partage de documents utilisant plusieurs alphabets. La transition vers une utilisation permanente d'Unicode, souvent à l'aide du codage UTF-8, est un processus d'envergure qui aboutira à terme à une gestion transparente des alphabets utilisés dans les divers documents échangés à l'intérieur de la société de l'information. Ce processus est également à l'œuvre dans des contextes fortement multilingues mais relativement peu informatisés : dans cet article, nous nous intéresserons au cas des langues africaines, à partir des observations recueillies lors de deux missions de formation au Niger, en 2000 et 2001.

Le Réseau international francophone d'aménagement linguistique (Rifal) visait dans son programme de formation 2000-2001 le soutien au traitement informatique de la langue française et des langues partenaires, à savoir les langues utilisées dans les pays du Sud, surtout africains, membres du Rifal : R.D. du Congo, Centrafrique, Madagascar, Niger, Sénégal. L'un des objectifs des sessions 2000 et 2001 était la mise en œuvre de techniques permettant la diffusion et le partage des ressources linguistiques grâce à l'utilisation d'Unicode et du réseau internet, en vue de la constitution de banques de données textuelles et terminologiques. Les deux missions de formation à Niamey (Niger) que nous avons contribué à assurer présentent un intérêt particulier dans la mesure où le Niger reconnaît huit langues nationales, à côté du français comme langue officielle. Toutefois, malgré les besoins que cette diversité engendre, les efforts d'informatisation des ressources linguistiques sont encore très limités dans cette partie de l'Afrique subsaharienne.

Nous allons décrire, dans ce qui suit, les défis posés par le multilinguisme à l'édition électronique, en les situant d'abord dans la perspective de l'alphabétisation des langues et des outils informatiques typiquement utilisés (section 2). Nous donnerons alors un aperçu du processus d'édition numérique observé dans un institut nigérien et de ses difficultés (section 3). Puis nous montrerons en quoi le standard Unicode permet de répondre à celles-ci (section 4), avant de fournir plusieurs perspectives sur un processus d'édition plus efficace, fondé sur Unicode (section 5).

2. Les défis de l'édition multilingue en Afrique

2.1. *Langues et écritures*

L'accès d'une langue à l'imprimerie, tout autant que l'accès au support numérique – deux révolutions souvent comparées – dépendent de l'existence d'une écriture offrant un support à cette langue. Plusieurs systèmes d'écriture sont apparus au cours de l'histoire, fondés sur un ensemble plus ou moins grand de caractères,

organisés en phrases sous une forme généralement linéaire. On classifie souvent les systèmes d'écriture en : (1) systèmes alphabétiques, fondés sur un petit nombre de signes à valeur souvent phonétique, tel l'alphabet latin ; (2) systèmes syllabiques, tels le syllabaire inuit ou le syllabaire japonais kana ; (3) systèmes idéographiques, exemplifiés de nos jours par les idéogrammes chinois, pour lesquels le nombre de signes avoisine le nombre de mots plutôt que celui de phonèmes ou de syllabes (plusieurs dizaines de milliers). Dans ce qui suit, nous nous intéresserons principalement aux systèmes alphabétiques, pour une raison facile à comprendre, liée à l'historique des langues africaines et de leur écriture. En effet, la plupart des systèmes idéographiques sont le fruit d'une évolution commencée il y a plusieurs siècles (ou dizaines de siècles), alors que les langues qui ont acquis plus récemment une écriture l'ont fait à partir d'un *alphabet* existant.

En Afrique, où certaines théories situent l'émergence évolutive du langage lui-même, le foisonnement des langues n'a pas entraîné le développement archaïque de systèmes d'écriture. Pour ce qui est des familles de langues (cf. Perrot, 1981, 1989), la diversité au sud du Sahara est telle que les linguistes ne sont pas encore parvenus à un accord sur le classement des quelques centaines de langues recensées (mais souvent peu étudiées). On distingue souvent les langues ouest-atlantiques (telles le peul et le wolof), les langues mandé plus au centre du continent, les langues voltaïques, celles du golfe de Guinée, enfin plus au sud la famille très étendue des langues bantoues, et la famille Khoisan. Revenant à l'équateur, on peut mentionner l'inclassable haoussa, ainsi qu'une importante zone de contact, riche en langues, entre le Nigeria et le Kenya. Ce panorama extrêmement sommaire laisse pour la fin les zones du nord et de l'est du continent, où un certain nombre de langues, plus facilement classables, ont développé des cultures écrites : les langues berbères, le domaine arabe (introduit à partir du VII^{e} siècle), l'ancien égyptien et le copte, le domaine éthiopien.

Si l'on excepte les hiéroglyphes de l'Egypte pharaonique, on rencontre toujours au nord et à l'est de l'Afrique des écritures alphabétiques, dont l'histoire peut remonter à l'Antiquité, tel l'alphabet *tifinagh* des langues berbères, dérivé du punique, ou l'écriture de l'amharique en Ethiopie (Christin, 2001). Toutefois, l'alphabétisation des langues africaines subsahariennes ne s'est produite qu'à une époque récente (XIX^{e}-XX^{e} siècles). Elle fait alors le plus souvent appel à l'alphabet latin, enrichi de nombreux caractères phonétiques modernes empruntés à l'Association internationale de phonétique, dans la mesure où des experts occidentaux sont le plus souvent à l'origine du processus. Un exemple de telles additions figure en annexe (section 7) pour une langue berbère du Niger. Mais historiquement d'autres alphabets ont pu servir de base pour l'alphabétisation, notamment l'alphabet arabe, actuellement utilisé au Niger pour un dialecte du *songhaï* nommé *tagdal*.

2.2. *L'informatisation des langues*

L'informatisation d'une langue peut se définir comme la possibilité de manipuler des documents stockés sous forme alphabétique, syllabique ou idéographique sur un support informatique, c'est-à-dire sous la forme d'une suite de caractères qui permet la manipulation du contenu linguistique. Une fois ce premier stade atteint, on peut évoquer l'existence d'outils informatiques adaptés à une langue, par exemple les outils d'édition qui facilitent la manipulation des documents, et aussi l'existence des ressources linguistiques : correcteurs d'orthographe, de grammaire, dictionnaires électroniques. On peut ensuite penser à la diffusion d'une langue sur internet, manifestée par le nombre de documents consultables dans des archives numériques ou indexés par des moteurs de recherche.

Les problèmes posés par le codage informatique des systèmes alphabétiques, syllabiques ou idéographiques diffèrent par leur ordre de grandeur entre les deux premiers types de systèmes et le dernier. La représentation informatique de l'alphabet latin dans sa version anglaise, contemporaine des premiers ordinateurs de série, a précédé de plusieurs décennies la représentation des trois principales écritures idéographiques, celle de la Chine, du Japon et de la Corée (système CJC). La dominance de l'alphabet latin en informatique est visible jusque dans la nature du codage informatique, qui donne toujours par défaut la priorité aux lettres latines utilisées en anglais.

2.3. *Les langues nigériennes et leur écriture*

Les six plus importantes langues nationales du Niger ont été abordées lors des formations Rifal 2000 et 2001 à Niamey. Il peut être utile de situer brièvement ces langues, toutes issues de familles linguistiques différentes (cf. Grimes et Grimes, 2000). Le *haoussa* est l'une des langues les plus parlées d'Afrique noire, avec environ 35 millions de locuteurs, dont 5 millions au Niger (plus de la moitié de la population). Le *zarma* est la langue de l'ethnie homonyme parlée par environ 1,5 millions de personnes au Niger, surtout dans la capitale. Le *peul* (ou *fulfulde*) totalise environ 15 millions de locuteurs en Afrique de l'Ouest, dont 800 000 au Niger. Le nombre de locuteurs du *tamachek* au Niger est aussi d'environ 800 000. Le *kanouri* est parlé par environ 4 % de la population du Niger, dans le dialecte *manga*. Enfin, le *gourmantché* est la moins représentée des langues abordées, avec environ 30 000 locuteurs au Niger (mais plus de 500 000 au Burkina Faso voisin). Les trois ou quatre groupes ethniques nigériens restants, chacun avec sa langue, ne dépassent pas 1 % de la population chacun.

Parmi les six langues abordées, le peul a été la moins étudiée : il semble que le jeu de caractères latins standard (ISO-latin-1, cf. ci-dessous), enrichi parfois de deux ou trois caractères supplémentaires, soit suffisant pour représenter son alphabet, du moins dans une version simplifiée. En revanche, les cinq autres langues utilisent à

des degrés divers des caractères spéciaux (entre quatre et seize), en général dérivés de l'alphabet latin. En effet, la définition d'une écriture pour ces langues s'est faite sur une base essentiellement phonétique, en utilisant les caractères de l'Association internationale de phonétique, basés sur l'alphabet latin.

Toutefois, la normalisation des écritures ne semble pas totalement achevée. Nos informations proviennent essentiellement des participants aux formations Rifal, tous de niveau universitaire. Il existe ainsi une norme pour l'écriture du haoussa, définissant les lettres et les groupes de lettres utilisés, ainsi qu'un ordre lexicographique (a, b, ɓ, c, d, ɗ, e, f, fy, g, gw, gy, h…). Pour le zarma et le tamachek, il existe des conventions officielles élaborées lors de colloques scientifiques, mais elles ne semblent pas universellement connues. Nous reviendrons plus bas sur le tamachek, mais signalons par exemple pour le zarma un usage encore incertain des caractères *i*, *e*, *o*, *u*, munis d'un tilde (tous rares en réalité).

La possibilité d'utiliser ces caractères spéciaux dans des programmes de traitement de texte n'est apparue que graduellement, en fonction des systèmes utilisés. Le logiciel TEX (puis LATEX), un des premiers et des plus robustes logiciels de publication par ordinateur, a fourni dès le début des années 1990 des polices munies des caractères spéciaux nécessaires aux langues européennes, puis à certaines langues africaines. Les propositions, par exemple celle de (Knappen, 1991) n'ont pu toutefois aboutir à la standardisation ; la police élaborée par ce dernier couvre *presque* tous les caractères utilisés au Niger, sauf ceux du tamachek (cf. notre Annexe), qui provient d'une famille linguistique non couverte par l'article. Signalons toutefois que le kanouri semble incomplètement couvert (bien que la langue soit citée explicitement), alors que le gourmantché et le zarma sont couverts sans être cités (c'est le songhaï, parent du zarma, qui est cité). Un autre logiciel, Sil Encore Font, permet de créer des polices de caractères personnalisées utilisables sur des ordinateurs personnels (PC et Apple). Muni d'une base de caractères essentiellement phonétiques, ce logiciel a été très utilisé par les linguistes de terrain, sans que des polices standardisées n'émergent toutefois.

Dans l'ensemble, la présence des langues nigériennes sur internet semble tout à fait occasionnelle, bien que nous ne disposions pas d'une étude approfondie. Par exemple, il existe une version de la *Déclaration universelle des droits de l'homme* en kanouri (dialecte *yerwa*) sur le site des Nations unies, qui utilise Unicode/UTF-8 (mais avec une police imposée par le site). Autre exemple, on trouve une version du *Pater* en haoussa, donnée simplement par le cliché de la page imprimée correspondante. Ces deux exemples montrent bien que la présence des langues sur internet est conditionnée par la possibilité d'encoder leurs alphabets.

3. L'édition de documents multilingues au Niger

3.1. *Vue d'ensemble : réalités et besoins*

Le développement de documents numériques dans les langues nationales du Niger en est seulement à ses débuts. La faible diffusion des technologies permettant la création et l'utilisation de tels documents n'est que l'une des raisons de cette situation – l'autre étant la présence du français comme langue officielle, de communication, et de culture. De ce fait, même les publications imprimées disponibles dans les langues nationales sont peu diversifiées, et leur qualité éditoriale laisse souvent à désirer. On constate en particulier que les polices de caractères utilisées ne permettent pas toujours d'écrire correctement la langue du document, mais sont choisies en fonction des capacités des éditeurs. La qualité des documents imprimés disponibles en français et en arabe est presque toujours supérieure à celle des documents dans les langues nationales.

Pourtant, le Niger déploie nombre d'efforts en direction des langues nationales, souvent de concert avec des organisations non gouvernementales. Un des domaines où les besoins sont les plus marquants est l'éducation, puisqu'il est établi que l'éducation de base en langue maternelle est bien plus efficace qu'une éducation en français. Or, les supports pédagogiques rédigés dans les langues nationales sont rares, et leur qualité va nécessairement de pair avec la capacité à représenter correctement les alphabets correspondants. Comme nous allons le voir, l'édition de ces documents étant en grande partie assistée par ordinateur, l'informatisation des langues nationales apparaît ainsi clairement comme prioritaire.

3.2. *La mission éditoriale de l'Indrap*

Les formations Rifal 2000 et 2001 ont été accueillies à Niamey par l'Indrap (Institut national de documentation, de recherche et d'animation pédagogiques). La mission première de l'institut est l'établissement des programmes d'enseignement pour les écoles nigériennes, accompagnés d'outils pédagogiques : manuels scolaires, guides pour enseignants, méthodes d'évaluation, émissions de radio pour enseignants. Les différentes « cellules » de l'institut regroupent des locuteurs des différentes langues nationales, mais la langue de communication et celle des documents de synthèse est le français. Du point de vue multilingue, l'une des principales difficultés est la traduction dans les langues nationales des termes français propres à chaque discipline d'enseignement. Cette tâche, indispensable pour la rédaction des programmes et manuels, échoit en général à la cellule de l'Indrap dédiée spécifiquement aux langues nationales. Enfin, la cellule de publication assistée par ordinateur permet à l'Indrap d'éditer de façon autonome l'ensemble des documents élaborés.

3.3. *L'édition numérique à l'Indrap : processus et documents*

Les nouvelles technologies ont changé bien des aspects de l'édition des documents, en particulier à l'Indrap. On assiste toutefois à un emploi particulier de ces technologies, conduisant à un processus d'édition original, qui diffère à la fois de l'imprimerie traditionnelle et de la publication assistée par ordinateur pratiquée en Occident. La *cellule de saisie et publication assistée par ordinateur (PAO)* de l'Indrap est en effet chargée de toutes les opérations de rédaction, à partir de la saisie des documents et jusqu'à la livraison à l'imprimeur des pages définitives. Les publications font de nombreux allers et retours entre les auteurs – qui fournissent initialement un manuscrit puisqu'ils ne disposent souvent pas d'ordinateurs – et la cellule, qui utilise des traitements de texte courants pour la mise en page des documents, et imprime à chaque stade une version sur papier. Ce procédé est donc relativement laborieux, même s'il faut reconnaître qu'il aboutit à une concentration des documents informatiques pouvant servir d'archives numériques pour l'Indrap.

Il paraît donc prématuré de parler d'une véritable utilisation des documents numériques en tant que tels. Pourtant, la mission pédagogique de l'Indrap, et la concentration d'utilisateurs potentiels, bénéficierait grandement d'un partage informatique de ces documents. Parmi ceux-ci, l'on a déjà cité les manuels, programmes, et outils d'enseignement ; ajoutons des documents ayant trait au multilinguisme, tels des lexiques bilingues des mathématiques ou des élections, traductions de documents officiels vers les langues nationales, textes littéraires.

Le principal problème posé à l'Indrap par le processus actuel, qui risque de s'aggraver avec l'augmentation du parc informatique, est la difficulté à représenter de manière uniforme les caractères spéciaux présents dans les différentes langues nationales. Les auteurs et les éditeurs de documents utilisent environ dix polices de caractères (certaines assez semblables), qui ont été définies sur place, souvent sans souci de standardisation. Dans la mesure où ces polices sont en général incompatibles, et qu'elles ne permettent pas la gestion de documents multilingues, souvent même excluant le français, le partage des documents est particulièrement difficile, seule la cellule PAO étant capable de trouver les polices adaptées.

4. Utilisation d'Unicode pour le codage informatique des langues africaines

4.1. *La place d'Unicode dans l'univers des polices de caractères*

Afin de faire comprendre au non spécialiste l'apport essentiel du standard Unicode aux problèmes décrits, il est utile de rappeler la situation en matière de codage de caractères avant l'apparition des outils permettant l'utilisation de ce standard. Cette situation caractérise encore bien le processus d'édition dans une institution comme l'Indrap au Niger.

De façon générale, les caractères d'un fichier informatique (un document) sont codifiés par une suite d'octets, un octet pouvant prendre une valeur entre 0 et 255. Jusqu'à l'avènement des outils mettant en œuvre le standard Unicode, les fichiers étaient lus octet par octet, et une interface graphique associait à chaque octet le caractère à afficher ou *glyphe* qui était prévu par la police de caractères utilisée.

Il faut distinguer ici la *police*, qui définit les formes exactes des caractères (par exemple, avec ou sans empattement), et le *jeu de caractères*, qui est une table de correspondances entre des nombres et des définitions abstraites de caractères (par exemple, « e latin avec accent aigu »). Ainsi, il existe plusieurs jeux de caractères standard à base latine, et pour chacun il existe de nombreuses polices, variant considérablement les formes des lettres. La dominance de l'alphabet latin en informatique se voit dans l'historique de ces jeux de caractères : un des premiers standards, ASCII, codait l'alphabet latin sur sept bits (le huitième servait de contrôle pour repérer les erreurs de transmission). Puis, libérant ce huitième bit, on a créé un jeu de 256 caractères, dont les emplacements nouveaux ont été attribués par différents standards à différentes combinaisons de caractères européens, le standard ISO-latin-1 (Europe occidentale) étant le plus utilisé. Toutefois, la création de plusieurs jeux de caractères à base latine, regroupés dans les standards ISO-latin de 1 à 10, procède de l'impossibilité d'ajouter à la base ASCII commune plus de 96 caractères différents en même temps (codes de 160 à 255). Les documents rédigés selon l'un de ces jeux sont ainsi très peu lisibles avec un autre jeu.

Afin d'unifier les très nombreux jeux de caractères existant au début des années 1990 (plusieurs jeux pour chaque écriture), il était devenu nécessaire d'utiliser une table de caractères sur plus d'un octet – et deux octets offrent déjà 65 536 places différentes. L'effort conjoint du consortium Unicode et de l'Organisation internationale de normalisation (ISO/CEI 10646) a abouti ainsi à un standard sur deux octets, qui fixe un code unique pour la plupart des caractères alphabétiques existants (dans tous les alphabets), pour de nombreux caractères symboliques (par exemple mathématiques), et pour des dizaines de milliers d'idéogrammes asiatiques.

La version 3.2 du standard Unicode, la plus récente, définit plus de 95 000 caractères, qui sont groupés par blocs pour en faciliter la description. Ainsi, la plupart des alphabets jouissent d'un bloc à part, par exemple le grec, le cyrillique, l'arménien, etc. Le tout premier bloc (U+0000 à U+0099 ou 0-255 en décimal), et le bloc Latin Etendu-A (U+0100 à U+017F), permettent d'écrire correctement la plupart des langues européennes. Il n'existe toutefois pas de bloc spécifiquement destiné aux langues africaines : on doit alors emprunter des lettres à différents blocs, notamment Latin Etendu-B (U+0180 à U+023F, prévu pour la transcription latine des alphabets non latins), et API (U+0250 à U+02AF, Alphabet phonétique international). Tout au plus le standard Unicode marque-t-il certaines lettres comme « africaines » ou « pan-nigérianes », mais cette sélection semble incomplète au vu de notre Annexe pour le tamachek (par exemple U+01CE ou U+1E6D devraient être marqués aussi). Il est vrai toutefois que la notion d'alphabet « africain » est

encore vague au niveau linguistique, ce qui explique son absence du standard Unicode.

Revenant à un plan plus technique, à la correspondance abstraite définie par Unicode, se sont ajoutés des codages en machine pour les caractères sur deux octets, qui sont des formes de sérialisation des bits d'information en vue du stockage. Le codage le plus répandu est UTF-8, avec un code de longueur variable (parfois plus de deux octets réels, souvent moins). En UTF-8, les 127 premiers caractères (le code ASCII) sont codés tels quels sur un octet, la plupart des caractères non idéographiques sont encodés sur deux octets, et le reste d'Unicode sur trois, puis quatre octets. UTF-8 assure ainsi que les documents ASCII existant déjà seront lus tels quels par un programme utilisant entièrement ce codage. Unicode/UTF-8 permet donc d'éditer des documents lisibles sur tout autre ordinateur, sans ambiguïté de police, car il n'est plus nécessaire d'utiliser des polices associées à tel ou tel jeu de caractères. Il faut seulement disposer d'une police compatible avec le standard Unicode, c'est-à-dire indexée selon les numéros de caractères Unicode. Ces polices ne diffèrent entre elles que par leur style et par le spectre de caractères Unicode disponibles. Parmi les polices avec plus de mille caractères contenant les caractères spéciaux qui nous intéressent ici, on peut citer Arial Unicode MS, Cyberbit Bitstream, Lucida Sans Unicode (la plus compacte).

4.2. *Unicode et le processus d'édition multilingue à l'Indrap*

Dans la perspective des polices/jeux de caractères, on peut décrire la situation initiale de l'Indrap dans les termes suivants. Grâce à différents outils de création de polices, l'institut dispose d'une dizaine de polices créées à partir de polices ISO-latin-1, mais dans lesquels bon nombre de codes, parfois inférieurs à 127 (donc ASCII), ont été remplacés par des glyphes correspondant aux besoins des langues nationales du Niger. En conséquence de ces redéfinitions, dont on comprend certes l'utilité, les caractères sont définis dans un document seulement à travers une police réalisant la correspondance entre le code (sur un octet) et le glyphe. Pour qu'un tel document rédigé sur un ordinateur soit lisible sur un autre, il faudra que ces polices créées localement y soient également installées. Ces polices locales sont donc un obstacle au partage des documents multilingues.

Dans la mesure où les alphabets des langues nationales nigériennes sont à base latine, mais utilisent entre cinq et seize caractères spéciaux (différents en général d'une langue à l'autre), on peut imaginer une solution en vue du partage des documents en ces langues, qui ne passe que par une police sur un octet (256 caractères) – comme nous l'esquisserons dans la section 5.2. Il est toutefois peu probable que le jeu de caractères implicite puisse devenir un standard. Il faudra donc toujours la fournir pour la visualisation des documents. En revanche, en utilisant le standard Unicode dès la rédaction des documents, leur partage sera immédiat et transparent (section 5.3). Pour l'heure, en vue de la conversion des

documents existants à Unicode (section 5.1), nous avons recensé à l'Indrap, pour chaque langue abordée, les caractères spéciaux, leurs codes dans chaque police, et les codes Unicode correspondants – cf. (Chanard et Popescu-Belis, 2000) et aussi les documents disponibles sur le site internet http://www.issco.unige.ch/osil/gtf-rifal mis en place par l'auteur.

4.3. *Etude d'un exemple : la langue tamachek*

Pour fournir une vision plus concrète, nous avons choisi de nous pencher sur l'une des langues abordées au Niger, le tamachek. Il s'agit d'une langue appartenant à la famille des langues berbères, parlée par les Touaregs du Sahara nigérien, et apparentée au kabyle d'Algérie. Outre l'alphabet *tifinagh*, le tamachek du Niger s'écrit aussi dans un alphabet à base latine, enrichi de plusieurs signes diacritiques et de deux caractères non latins – en tout, seize caractères spéciaux. Ce nombre n'est pas surprenant, puisque le tamachek est une langue sémitique, avec un système de consonnes bien plus riche que celui des langues latines.

Afin de produire des documents électroniques en tamachek, les chercheurs de l'Indrap ont conçu plusieurs polices de caractères dont le jeu de caractères implicite est fondé sur ISO-latin-1, dans lesquelles un certain nombre de caractères latins sont remplacés par des caractères propres au tamachek. Les polices ont été créées grâce au paquetage SIL Encore Font, diffusé lors de précédentes formations Rifal.

L'analyse minutieuse que nous avons entreprise montre que les polices locales sont souvent incompatibles entre elles, et souvent incomplètes. Le tableau donné en Annexe (section 7) répertorie les caractères spéciaux du tamachek, et leur code dans les polices étudiées (les polices sont désignées seulement par une lettre majuscule). On voit que seules les polices I et T contiennent tous les caractères nécessaires. Les codes (octets) utilisés pour le tamachek sont en général convenables, surtout pour la police T, mais parfois des caractères importants du jeu ISO-latin-1 sont malencontreusement redéfinis. Par exemple, T redéfinit le « p majuscule » en « schwa majuscule » (code 80), et I redéfinit le « @ » en « s minuscule avec point souscrit » (code 64) – cela parmi d'autres redéfinitions problématiques signalées par le caractère « ! » dans notre tableau. Enfin, dans la police I, deux caractères sont définis deux fois. L'existence de deux polices particulières au tamachek à l'intérieur même de l'Indrap fait obstacle à la communication des documents électroniques.

5. Vers l'édition et le partage des documents numériques multilingues

5.1. *Solution de transition : partage grâce au format HTML*

Une première solution d'interopérabilité a été proposée dans le cadre des formations Rifal 2000-2001. Elle consiste à convertir les documents que l'on

souhaite partager vers le format HTML, c'est-à-dire vers des fichiers hypertexte pourvus de balises, du type rencontré sur internet. Naturellement, il faut prendre soin d'encoder les caractères spéciaux propres à chaque langue par les entités HTML de type « &#XYZ; » ayant un code hexadécimal Unicode « XYZ » correspondant bien au caractère spécial à afficher. Il faut aussi préciser en tête du document que son codage est bien Unicode. On peut également, et cela est plus compact, introduire directement dans le fichier le code UTF-8 correspondant au code Unicode « XYZ » (de un à quatre octets), et préciser ensuite la forme de sérialisation, à savoir UTF-8.

De cette façon, les documents sont lisibles par n'importe quel programme sachant afficher des documents HTML (navigateurs internet, éditeur MS Word, etc.), et les caractères spéciaux sont affichés correctement du moment que ce programme comprend les entités HTML/Unicode et qu'il possède une police possédant le sous-ensemble approprié de caractères Unicode (à base latine). De nombreuses polices peuvent convenir, du moment où elles sont compatibles avec les numéros Unicode – par exemple Lucida Sans Unicode.

Afin d'implémenter la solution proposée, deux étapes sont distinguées dans la conversion des documents existants. La première consiste à convertir les documents depuis leurs formats respectifs vers HTML, grâce aux logiciels d'édition les ayant créés (le plus souvent MS Word). Ces logiciels ne connaissent pas la correspondance entre la police locale utilisée et Unicode, c'est pourquoi ils utilisent des balises HTML de type <FONT> pour signaler l'utilisation de la police respective, et maintiennent le code sur un octet utilisé à l'origine. La seconde étape consiste à supprimer les balises <FONT> et à remplacer les codes sur un octet soit par les entités Unicode, soit (plus compact) par les codes UTF-8 correspondants. Pour ce faire, un logiciel intitulé Conv2utf8 a été utilisé dans les formations Rifal. Ce logiciel a été conçu et réalisé par C. Chanard au LLACAN-CNRS (Villejuif, France) et est décrit plus en détail dans (Chanard et Popescu-Belis, 2002). Le logiciel parcourt le fichier et effectue les remplacements appropriés grâce à des tables de correspondances entre les polices locales et Unicode.

Ce processus possède quelques inconvénients, mais surtout offre de nombreux avantages. Il y a d'abord la difficulté de définir les tables de correspondances, mais ce travail ne doit être fait qu'une seule fois, et nous l'avons déjà réalisé pour la dizaine de polices de l'Indrap. On peut aussi avancer que HTML n'est pas un format d'édition très pratique, mais en réalité le procédé est surtout orienté vers le partage des documents en vue de leur visualisation. Enfin, il est important de noter que HTML est un format non propriétaire, transparent, universellement compris, donc bien adapté au partage des documents numériques multilingues. L'utilisation du standard XML pour le stockage des documents, et des feuilles de style XSL (XSLT ou XSL-FO) pour le formatage en vue de l'affichage est une solution d'inspiration analogue, plus puissante mais plus complexe ; elle est à l'étude pour les prochaines formations du Rifal (2002-2004).

L'avantage principal du procédé proposé est de permettre la conservation et l'archivage durable des documents existants, ainsi que leur partage sans perte d'information à travers un réseau local ou à travers internet. Fruit des formations Rifal 2000-2001, une archive électronique contenant de tels documents est en cours de constitution (voir le site http://www.issco.unige.ch/osil/gtf-rifal/).

5.2. *Définition d'une police commune sur un octet*

Une solution permettant l'édition multilingue et en même temps le partage des documents en format d'édition (par exemple MS Word) est la constitution d'une police sur un octet (0-255) qui soit adoptée par tous les éditeurs de documents de l'Indrap, ou mieux encore, du Niger. Une telle police ne permettrait certes pas de visualiser correctement les documents existants, puisqu'ils ont été créés avec des polices qui ne pourront pas être toutes compatibles avec le nouveau standard. En ce sens, la solution précédente garde toute sa valeur.

Afin de créer une telle police « nigérienne », il faut tout d'abord déterminer les codes, dans une police de 256 caractères, qui peuvent être redéfinis sans trop affecter les codes du standard ISO-latin-1, en particulier les caractères accentués du français. Cette analyse a été fournie aux partenaires du Rifal et figure aussi dans (Chanard et Popescu-Belis 2002). Il en ressort qu'environ 60 codes (tous supérieurs à 128) sont redéfinissables, ce qui permet de définir convenablement des codes pour les caractères des langues nigériennes (majuscules et minuscules). Par opposition, il faut savoir qu'aucune des polices utilisées actuellement n'observe entièrement ces recommandations. Parallèlement, un système de raccourcis clavier devra être mis en place, et des efforts conjoints seraient nécessaires pour parvenir à remplacer toutes les polices précédemment utilisées.

5.3. *Un processus d'édition multilingue fondé sur Unicode*

L'enrichissement du parc informatique de l'Indrap et le développement d'un réseau local aideraient certainement à améliorer le processus d'édition multilingue, en diminuant en particulier le nombre d'allers et retours des documents entre les auteurs et les éditeurs. Outre le fait que la tâche d'édition serait alors distribuée, on pourra aussi munir les postes de travail individuels d'outils logiciels permettant d'utiliser d'emblée le standard Unicode, dès le processus de rédaction. Il devient en effet visible que les principaux outils d'édition deviennent capables de manipuler l'ensemble du jeu de caractères Unicode, ce qui diminue l'intérêt des polices locales décrites ci-dessus (leur intérêt demeure si l'on souhaite varier la forme des glyphes). Toutefois, il serait dangereux de s'enfermer complètement dans un format de données propriétaire, et à ce titre la première solution proposée permet mieux le partage transparent des documents. Dans la situation décrite ici, le logiciel Conv2utf8 deviendrait superflu, puisque les outils d'édition Unicode génèrent des

fichiers HTML directement en UTF-8. Le problème de la saisie des caractères spéciaux est quant à lui reporté sur le logiciel utilisant Unicode – un système de raccourcis clavier est probablement la solution.

L'ensemble de ces améliorations, qui vont de l'utilisation de HTML/UTF-8 comme format pivot jusqu'aux outils d'édition intégrant directement Unicode, sont en mesure d'aider à la constitution d'une base de textes et d'une banque de données terminologique, multilingues, au Niger. Il s'agit d'une perspective prometteuse pour de nouvelles actions du Rifal en direction des langues africaines, en priorité dans le domaine de la pédagogie des différentes disciplines d'enseignement. En effet, de telles ressources permettraient de répondre à des besoins réels en matière d'instruction scolaire au Niger, mais seraient aussi un instrument de recherche précieux pour les linguistes africanistes. Elles permettraient également de consolider le rôle du français dans l'aménagement linguistique en Afrique subsaharienne, ce qui correspond à la mission du Rifal dans le cadre de la Francophonie.

Remerciements

L'auteur souhaite remercier Christian Chanard (LLACAN-CNRS, Paris) pour sa collaboration. L'auteur est reconnaissant pour leur soutien aux formations Rifal 2000-2001 au Niger à : Marcel Diki-Kidiri (LLACAN-CNRS et GTF/Rifal, Paris), Marcel Grangier (Chancellerie fédérale, Berne), et Louis-Jean Rousseau (Rifal et Office de la langue française, Québec). L'auteur remercie également les professeurs Bruno de Bessé et Margaret King (Université de Genève), ainsi que le directeur général de l'Indrap, Djibo Seybou Kalilou. Les remarques de Don Osborn (Bisharat.net) et Andrew Cunningham (State Library of Victoria, Australie) ont été utiles à cet article, ainsi que celles de Jacques André (IRISA, Rennes) et de Patrick Andries (Montréal). Enfin, il a été un plaisir de coencadrer les formations Rifal au Niger avec Florian Simmen (Université de Genève), puis avec Marcel Diki-Kidiri.

6. Bibliographie

Chanard C., Popescu-Belis A., « Encodage informatique multilingue : application au contexte du Niger », *Cahiers du RIFAL*, n° 22, p. 33-45, 2002.

Christin, A.-M., *Histoire de l'écriture*, Paris, Flammarion, 2001.

Consortium Unicode, *The Unicode Standard Version 3.0*, Redding, MA, Addison-Wesley, 2000. Voir aussi www.unicode.org.

Czyborra R., The global character set Unicode in the Unix operating system, Mémoire de diplôme, Technische Universität Berlin, 1998. Voir aussi www.czyborra.com.

Grimes B.F., Grimes J.E., éds., *Ethnologue 14th Edition*, Dallas TX, SIL International Academic Bookstore, 2000. Voir aussi www.sil.org/ethnologue/

ISO/CEI 10646-1 :2000, *Technologies de l'information – Jeu universel de caractères codés sur plusieurs octets (JUC) – Partie 1 : Architecture et plan multilingue de base*, Genève, Organisation Internationale de Normalisation, 2000. Voir aussi www.iso.ch.

Knappen, J., « TeX in Africa », *Cahiers GUTenberg*, n° 10-11, (*Actes de la 6ème conférence TeX européenne & GUTenberg'91*, Paris, 23-26 septembre 1991), 1991, p. 15-24. Voir aussi www.gutenberg.eu.org/pub/GUTenberg/publications/cahiers.html.

Perrot, J., éd., *Les langues dans le monde ancien et moderne*, vol. I, *Les langues de l'Afrique subsaharienne – Pidgins et créoles*, Paris, Editions du C.N.R.S., 1981.

Perrot, J., éd., *Les langues dans le monde ancien et moderne*, vol. II, *Les langues chamito-sémitiques*, Paris, Editions du C.N.R.S., 1989.

Organisation des Nations unies, « Déclaration universelle des droits de l'homme », *Résolution l'Assemblée générale*, 217 A (III), 10.12.1948. Voir aussi www.unhchr.ch/udhr/index.htm.

The Unicode Consortium, *The Unicode Standard, Version 3.0*, Reading, MA, Addison-Wesley, 2000. Voir aussi en français hapax.iquebec.com.

7. Annexe

Glyphe	Code Unicode		SILID	Code dans la police					
	hexa.	déc.		A	I	L	N	H	T
ă	01CE	0462	0097 +6006	-	210	~ 181	-	-	228
Ă	01CD	0461	0065 +6106	-	! 145 ! 189	-	-	-	196
ḍ	1E0D	7693	0100 +6613	-	! 151	-	-	-	230
Ḍ	1E0C	7692	0068 +6613	-	! 190	-	-	-	198
ə	01DD	0477	1106	!! 038 !! 142	168	176	! 034	-	!! 112
Ǝ	018E	0398	2106	163	! 224	223	! 035	-	!! 080
ḷ	1E37	7735	0108 +6613	-	! 179	-	-	-	242
Ḷ	1E36	7734	0076 +6613	-	176	-	-	-	210
ṣ	1E63	7779	0115 +6613	-	!! 064	-	-	-	156
Ṣ	1E62	7778	0053 +6613	-	216	-	-	-	140 ! 189
š	0161	0353	7023	154	154	!! 153	154	154	! 190
Š	0160	0352	8023	138	! 137	! 137	138	138	191
ṭ	1E6D	7789	0116 +6613	-	!! 182	-	-	-	! 253
Ṭ	1E6C	7788	0084 +6613	-	227 !! 036	-	-	-	! 221
ẓ	1E93	7827	0122 +6713	-	184	-	-	-	255
Ẓ	1E92	7826	0090 +6613	-	!!153	-	-	-	159
ɣ	0263	0611	1429	-	177	128	-	-	163
ɣ	0194	0404	-	-	-	-	-	-	-
⊗	2297	8855	-	-	-	-	-	-	-

Figure 1. *Codes des différents caractères spéciaux du tamachek dans les différentes polices testées à l'Indrap (A, I, L, N, H, T). L'absence de code est notée par « - », et un code mal choisi est noté par « ! » ou « !! » en fonction de la gravité de l'erreur*