Vers l'informatisation de quelques langues d'Afrique de l'Ouest

Chantal Enguehard[1] *Soumana Kané*[2] *Mathieu Mangeot*[3] *Issouf Modi*[4] *Mamadou Lamine Sanogo*[5]
(1) LINA2, rue de la Houssinière, BP 92208, 44322 Nantes Cedex 03, France
(2) CNR-ENF, BP 62, Bamako, Mali
(3) LIG,BP 53 38041 Grenoble, France
(4) MEN/A/PLN/DGPLN/DREL, BP 557, Niamey, Niger
(5) CNRST, BP 7047 Ouagadougou 03, Burkina Faso
chantal.enguehard@univ-nantes.fr, soumanak@yahoo.com, Mathieu.Mangeot@imag.fr, modyissouf@yahoo.fr, mala_sng@yahoo.fr

RÉSUMÉ

Le projet DILAF vise à établir une méthodologie de conversion de dictionnaires éditoriaux en des fichiers XML au format (Lexical Markup Framework) et à l'appliquer sur cinq dictionnaires. Nous présentons les motivations de ce projet puis les dictionnaires concernés ainsi que les alphabets des langues de ces dictionnaires. Il s'agit de dictionnaires bilingues langue africaine-français : haoussa-français, kanouri-français, soŋay zarma-français, tamajaq-français et bambara-français. La présentation de la plateforme jibiki de manipulation des ressources lexicales est suivie de l'exposé des travaux menés en collaboration avec les linguistes, informaticiens et lexicographes participant au projet. La cinquième partie établit un bilan quant à la représentation des caractères des différentes langues dans Unicode et détaille le cas particulier des caractères tifinagh. Les travaux futurs sont ensuite évoqués.

ABSTRACT

The DILAF project aims to establish a methodology to convert of editorial dictionaries into XML files expressed according with rhe LMF (Lexical Markup Framework) format and to apply tis mothodology on five dictionaries. We present the motivation of this project, then the concerned dictionaries and the alphabets of the languages of these dictionaries. These are bilingual dictionaries Africanlanguage-French: Hausa-French, Kanuri-French, Soŋay Zarma-French, Tamajaq-French and Bambara-French. The jibiki platform is presented, then we detail the adavances of the project thanks to the collaboration of linguists, computer scientists, and lexicographers. The fifth part establishes a balance concerning the Unicode representation of the characters of the different languages and details the particular case of the tifinagh characters.

MOTS-CLÉS : LMF, TALN, dictionnaires, langues africaines, Unicode

KEYWORDS : LMF, NLP, dictionnaries, African languages, Unicode

1 Motivation

Si l'accès aux ordinateurs est considéré comme le principal indicateur de la fracture numérique en Afrique, il faut reconnaître que la faible disponibilité des ressources dans les langues africaines constitue un handicap dont les conséquences sont incalculables pour le développement des Technologies de l'Information et de la Communication (TIC)

JEP-TALN-RECITAL 2012, Atelier TALAf 2012: Traitement Automatique des Langues Africaines, pages 27–40,
Grenoble, 4 au 8 juin 2012.

dans cette partie du monde. Aussi, la production, la diffusion et la vulgarisation de ressources locales adaptées dans ces langues nous paraissent-elles être indiquées pour une implantation durable des TIC sur le continent. Or, la plupart des langues de l'espace francophone d'Afrique de l'Ouest sont peu dotées (langues-pi) (Berment, 2004) : les ressources électroniques disponibles sont rares, mal distribuées, voire inexistantes. Seules sont accessibles les fonctions d'édition et d'impression rendant l'exploitation de ces langues difficile au moment où il est question de les introduire dans le système éducatif, de créer des normes d'écriture standardisées et stabilisées et surtout de développer leur usage à l'écrit dans l'administration et la vie quotidienne.

Aussi, afin de contribuer à combler ce retard, nous – collègues du Sud et du Nord – nous sommes engagés à améliorer l'équipement de quelques langues africaines à travers, entre autres, l'informatisation de dictionnaires éditoriaux portant sur des langues africaines. A cet effet, nous présenterons le projet DiLAF (Dictionnaires Langues Africaines Français) qui vise à convertir des dictionnaires éditoriaux bilingues en un format XML1[1] permettant leur pérennisation et leur partage (Streiter et al., 2006). Ce projet international rassemble des partenaires du Burkina Faso (Centre National de la Recherche Scientifique et Technologique), de France (Laboratoire d'Informatique de Grenoble et Laboratoire d'informatique de Nantes-Atlantique), du Mali (Centre National de Ressources de l'Éducation Non Formelle) et du Niger (Institut National de Documentation de Recherche et d'Animation Pédagogiques, Ministère de l'Education Nationale, et Université Abdou Moumouni de Niamey).

En nous fondant sur un travail de base déjà effectué par des lexicographes nous avons constitué des équipes pluridisciplinaires constituées de linguistes, d'informaticiens et de pédagogues. Cinq dictionnaires comportant, chacun, plusieurs milliers d'entrées, devraient être convertis et intégrés à une plate-forme Jibiki de gestion de ressources lexicales (Mangeot, 2001). Les dictionnaires seront donc disponibles sur Internet d'ici la fin de l'année 2012 sous licence Creative Commons.

— dictionnaire bambara-français, Charles Bailleul, édition 1996,
— dictionnaire haoussa-français destiné à l'enseignement du cycle de base 1, 2008, Soutéba,
— dictionnaire kanouri-français destiné pour le cycle de base 1, 2004, Soutéba,
— dictionnaire soŋay zarma-français destiné pour le cycle de base 1, 2007, Soutéba,
— dictionnaire tamajaq-français destiné à l'enseignement du cycle de base 1, 2007, Soutéba.

Il s'agit de dictionnaires d'usage qui visent surtout à vulgariser les formes écrites de l'usage quotidien des langues africaines dans la pure tradition lexicographique (Matoré, 1973), (Eluerd, 2000). Se démarquant des démarches normatives et dirigistes des dictionnaires normatifs (Mortureux, 1997), les présents dictionnaires descriptifs restent ouverts aux contributions et leur mise en ligne devra, nous l'espérons, développer un sentiment de fierté chez les usagers des différentes langues. De même, ils participeront au développement d'un environnement lettré propice à l'alphabétisation dont le faible taux compromet les acquis des progrès réalisés dans les autres secteurs.

Nous présenterons l'origine et la structure de ces dictionnaires ainsi que quelques

[1] Extended Markup Language.

entrées, puis les résultats de l'atelier de démarrage qui s'est déroulé du 6 au 17 décembre 2010 à Niamey (Niger). Ensuite nous détaillons les constats réalisées quant à la prise en compte de ces langues par le standard Unicode et par les logiciels que nous avoins utilisés. Enfin nous évoquons les futurs travaux.

2 Cinq dictionnaires bilingues langue africaine-français

Quatre des cinq dictionnaires sur lesquels nous travaillons ont été produits par le projet Soutéba (programme de soutien à l'éducation de base) avec le financement de la coopération allemande1[2] et l'appui de l'Union Européenne. Ces dictionnaires, destinés à l'éducation de base, sont de structure simple car ils ont été conçus pour des enfants de classe primaire scolarisés en école bilingue (l'enseignement y est donné en une langue nationale et en français). La plupart des termes de lexicologie, telles les étiquettes lexicales ou les catégories grammaticales, les signalisations de synonymies, d'antonymies, de genres, de variations dialectales, etc., y sont notés dans la langue dont il est question dans le dictionnaire, contribuant ainsi à forger et à diffuser un méta-langage dans la langue locale ainsi qu'une terminologie spécialisée. Les entrées sont énoncées en ordre alphabétique, même dans le cas du tamajaq (bien qu'il soit habituel de présenter les entrées de cette langue en fonction des racines) car les voyelles sont explicitement écrites (ce mode de classement a été privilégié car il est bien connu des enfants).

2.1 Dictionnaire haoussa-français

Il comprend 7823 entrées classées selon l'ordre lexicographique suivant : a b ɓ c d ɗ e f fy g gw gy h i j k kw ky ƙ ƙw ƙy l m n o p r s sh t ts u w y y̆ z (Arrêté, 212-99).

Elles sont structurées avec des schémas différents selon la catégorie grammaticale. Toutes les entrées sont d'ordre orthographique ; suivent la prononciation (les tons sont marqués par les signes diacritiques posés sur les voyelles) et la catégorie grammaticale. Sur le plan sémantique, il existe une définition en langue haoussa, un exemple d'emploi (repéré par l'usage de l'italique), puis l'équivalent en français. L'entrée d'un nom précise en sus le genre, le féminin s'il existe, le ou les pluriels (selon les genres) et les éventuelles variantes dialectales. Pour les verbes, il est parfois nécessaire de préciser les degrés pour calculer les dérivés morphologiques. Les variantes morpho-phonologiques des dérivations féminine et plurielle des adjectifs sont énoncées.

Exemple :

jaki [jàakíi] *s.* **babbar dabbar gida mai kamar doki, wadda ba ta kai tsawon doki ba amma ta fi shi dogayen kunnuwa**. ***Ya aza wa jaki kaya za ya tafi kasuwa.*** *Jin.:* n. *Sg.:* **jaka**. *Jam.:* **jakai, jakuna**. *Far.:* âne

Le mot "jaki" se prononce [jàakíi]. Sa catégorie grammaticale est "s.", abbréviation de "suna" qui signifie nom.

Sa définition est : "babbar dabbar gida mai kamar doki, wadda ba ta kai tsawon doki ba amma ta fi shi dogayen kunnuwa."

Un exemple d'usage est signalé en caractères italique : "Ya aza wa jaki kaya za ya tafi

[2] DED : Deutscher Entwicklungsdienst.

kasuwa."

"Jin.", abréviation de "jinsi" (genre) précède ici "n.", abréviation de "namiji" (masculin).

Plusieurs variations morphologiques sont signalées. La forme féminine "jaka" suit l'abréviationg.", les formes plurielles "jakai" et "jakuna" sont signalées par "Jam.", abbabréviation "jam'i" (pluriel). L'équivalent en français, signalé par "Far." ("faransanci"), clôt l'entrée.

2.2 Dictionnaire kanouri-français

Le dictionnaire kanouri-français comprend 5994 entrées classées selon l'ordre lexicographique suivant : a b c d e ə f g h i j k l m n ny o p r ɾ s sh t u w y z (Arrêté, 213-99).

La forme orthographique de l'entrée est suivie d'indications de prononciation ciblées sur la notation des tons : le ton haut est noté par un accent aigu, le ton bas par un accent grave, le ton montant par un caron (signe suggérant la succession d'un accent grave et d'un accent aigu) et le ton descendant par un accent circonflexe (signe suggérant la succession d'un aigu et d'un accent grave). La catégorie grammaticale de l'entrée est indiquée en italique. Une définition, un exemple d'usage puis le sens en français sont ensuite énoncés. D'autres informations peuvent apparaître comme des variantes.

Exemple :

abəɾwa [àbə̀ɾwà] *cu.* **Kəska təngəri, kalu ngəwua dawulan tada cakkidə.** ***Kəryende kannua nangaro, abəɾwa cakkiwawo.*** [*Fa.:* ananas]

Le mot "abəɾwa" se prononce [àbə̀ɾwà]. Sa catégorie grammaticale est "cu." (nom).

Sa définition est écrite en caractères gras : "Kəska təngəri, kalu ngəwua dawulan tada cakkidə."

Un exemple d'usage est signalé en caractères italique : "Kəryende kannua nangaro, abəɾwa cakkiwawo."

L'équivalent en français, précédé de "Fa.:" et encadré de crochets, termine l'entrée.

2.3 Dictionnaire soŋay zarma-français

Il comprend 6916 entrées classées selon l'ordre lexicographique suivant : a ã b c d e ẽ f g h i ĩ j k l m n ŋ ɲ o õ p r s t u ũ w y z (Arrêté, 215-99).

Chaque entrée présente une forme orthographique suivie d'une transcription phonétique dans laquelle les tons sont notés selon les conventions déjà exposées pour le kanouri (partie 1.2). La catégorie grammaticale précise explicitement, pour les verbes, la transitivité ou l'intransitivité. Pour certaines entrées, des antonymes, synonymes ou renvois sont indiqués. Une glose en français, une définition et un exemple terminent l'entrée.

Exemple :

ɲagas [ɲágás] *mteeb.* • *brusquement (détaler)* • sanniize no kaŋ ga cabe kaŋ boro na zuray sambu nda gaabi sahã-din • *Za zankey di hansu-kaaro no i te ɲagas*

Le mot "ɲagas" se prononce [ɲágás]. Sa catégorie grammaticale est "mteeb." (adverbe).

L'équivalent en français est signalé en caractères italiques.

Sa définition est : "sanniize no kaŋ ga cabe kaŋ boro na zuray sambu nda gaabi sahã-din"

Un exemple d'usage est énoncé en caractères italiques : "Za zankey di hansu-kaaro no i te ɲagas"

2.4 Dictionnaire tamajaq-français

Le dictionnaire tamajaq-français comprend 5205 entrées du parler təwəlləmmət classées selon l'ordre lexicographique suivant : a ă ə b c d ḍ e f g ǧ h i j ǰ ɣ k l ḷ m n ŋ o q r s ṣ š t ṭ u w x y z ẓ (Arrêté, 214-99)[3]. Les voyelles longues sont notées par un accent circonflexe : â, ê, û; ô, î.

La forme orthographique de l'entrée est suivie de la catégorie grammaticale de l'entrée et d'une glose en français indiquées en italique. Pour les noms figurent souvent des indications morphologiques concernant l'état d'annexion ; le pluriel et le genre sont souvent explicitement indiqués. Une définition, un exemple d'usage sont ensuite énoncés. D'autres informations peuvent apparaître comme des variantes, des synonymes, etc.. Le tamajaq n'étant pas une langue tonale, la phonétique n'apparaît pas.

Exemple :

əbeɣla *sn. mulet* ♦ **Ag-anɣer əd tabagawt. Ibeɣlan wər tăn-tăha tămalăɣa.** *anammelu.:* **făkr-ejăḍ.** *təmust.:* **yy.** *iget.:* **ibəɣlan**.

Le mot "əbeɣla" est un "sn.", abbréviation de "isən" (nom) qui signifie mulet en français.

Sa définition "Ag-anɣer əd tabagawt." et un exemple d'usage "Ibeɣlan wər tăn-tăha tămalăɣa." sont écrits en caractères gras.

Un synonyme (anammelu) est signalé : "făkr-ejăḍ".

Le genre (təmust) est "yy.", abréviation de "yey" (masculin).

Le pluriel de ce mot (iget) est "ibəɣlan".

2.5 Dictionnaire bambara-français

Le dictionnaire bambara-français du Père Charles Bailleul (édition 1996) comprend plus de 10 000 entrées ordonnées selon l'ordre lexicographique suivant : a b c d e ɛ f g h i j k l m n ɲ ŋ o ɔ p r s t u w y z.

Ce dictionnaire est d'abord destiné aux locuteurs français désireux de se perfectionner en bambara mais il constitue également une ressource pour les bambaraphones. Selon les dires de l'auteur lui-même, il « se veut être un outil de travail au service de l'alphabétisation, l'enseignement et la culture bambara ». A ce jour, il peut être considéré comme le dictionnaire le plus fourni et le plus complet sur cette langue. Aussi il est

[3] Les signes ǰ' et ' ǧ' sont utilisés uniquement pour transcrire certains parlers comme celui de l'Ayər, par conséquent ils n'apparaissent pas dans ce dictionnaire.

consulté par les spécialiste des autres variétés de cette langue que sont le dioula (Burkina Faso, Côte d'Ivoire) et le manlinké (Guinée, Gambie, Sierra Leone, Libéria, etc.).

Bien que l'orthographe du bambara ne note pas les tons, et ce par économie de signes, les tons sont marquées dans toutes les entrées et tous les exemples d'usage : l'accent grave sur une voyelle brève marque un ton bas ponctuel ("bìnɔ̀gɔ̀kɛ" – "oncle paternel") ; l'accent grave sur une voyelle répétée l'affecte sur toute sa longueur ("dɛ̀ɛmu" – "parole" – se prononce dɛ̀ɛ̀mu); l'accent grave suivi d'un accent aigu marque une voyelle longue relevée sur sa deuxième partie (ex : "ɲàá" – "nid)" ; le caron marque un ton bas modulé ascendant (ex : "bě" – "accord").

La prononciation phonétique n'est précisée que lorsque l'orthographe officielle s'écarte de la prononciation effective. Dans de tels cas, elle figure entre crochets. Par exemple, pour l'entrée « da.lan [dlan] (...) n. lit » l'indication phonétique [dlan] indique que "dalan" n'est jamais prononcé complètement, c'est-à-dire en deux syllabes.

Les entrées, surtout complexes, sont accompagnées de leur origine et de leur structure, car il s'agit d'informations nécessaires pour une bonne traduction. Ainsi, pour les dérivés et composés, l'analyse des éléments est indiquée entre parenthèses et la frontière sémantique suggérée par un point, comme dans l'entrée suivante : « ɲɛmɔgɔ ɲɛ.mɔgɔ (devant.personne) dirigeant, chef. […] » Cette présentation de l'entrée indique que, morphologiquement, "ɲɛmɔgɔ" se compose de "ɲɛ" et de "mɔgɔ" (ce qui est indiqué par le point) et que, sémantiquement, dans l'ordre, il signifie "devant" et "personne" (ce qui est indiqué par les parenthèses et le point), le sens de tout le composé se ramenant à dirigeant, c'est-à-dire une personne placée devant, à la tête de… (traduction privilégiée indiquée par le soulignement).

On peut ainsi multiplier les exemples :

« kalanso kàlàn.so (instruction.maison) classe d'école » : mot composé de "kalan" et "so", respectivement "instruction" et "maison", signifie "classe d'école".

« mɔgɔdun mɔgɔ.dun (personne.manger) cannibale, anthropophage » : mot composé de "mɔgɔ" et "dun", respectivement "personne" et "manger", signifie "cannibale".

« juguya jugu.ya (mauvais.suff abst) méchanceté » : mot dérivé ("jugu" et "-ya", respectivement "mauvais" et suffixe d'abstraction), signifie "méchanceté".

« walanba walan.ba (tablette.suff augm) tableau noir » : mot dérivé ("walan" et "-ba", respectivement "tablette" et suffixe augmentatif), signifie "tableau noir".

Il est important de signaler que la dérivation et la composition étant des procédés très productifs en bambara, les cas retenus pour figurer dans le dictionnaire ont été choisis en fonction de leur fréquence d'emploi et de leur variation de sens par rapport à leur formation.

L'origine des emprunts est indiquée entre accolades : {fr} pour le français, et {ar} pour l'arabe.

Exemples : « kaso kàso {fr: cachot} n. Prison » ; « ala ala {ar: allah = Dieu} »

Enfin, ce dictionnaire accorde quelque place aux néologismes proposés par les services d'alphabétisation. Il s'agit notamment de « ceux qui sont les plus utilisés ou semblent

promis à un bel avenir ». Ils sont signalés par l'indication (néologisme).

Exemples : « kumaden kuma.den (parole.élément) mot (néologisme) » ; « kɔbila kɔ.bila (derrière.placer) postposition (néologisme) »

3 Plate-forme jibiki

Jibiki (Mangeot et al., 2003; Mangeot et al., 2006) est une plate-forme générique en ligne pour manipuler des ressources lexicales avec gestion d'utilisateurs et groupes, consultation de ressources hétérogènes et édition générique d'articles de dictionnaires. Ce site Web communautaire a initialement été développé pour le projet Papillon (http://www.papillon-dictionary.org). La plate-forme est programmée entièrement en Java, fondée sur l'environnement "Enhydra". Toutes les données sont stockées au format XML dans une base de données (Postgres). Ce site Web propose principalement deux services : une interface unifiée permettant d'accéder simultanément à de nombreuses ressources hétérogènes (dictionnaires monolingues, dictionnaires bilingues, bases multilingues, etc.) et une interface d'édition spécifique pour contribuer directement aux dictionnaires disponibles sur la plate-forme.

L'éditeur (Mangeot et al., 2004) est fondé sur un modèle d'interface HTML instancié avec l'article à éditer. Le modèle peut être généré automatiquement depuis une description de la structure de l'entrée à l'aide d'un schéma XML. Il peut être modifié ensuite pour améliorer le rendu à l'écran. La seule information nécessaire à l'édition d'un article de dictionnaire est donc le schéma XML représentant la structure de cette entrée. Par conséquent, il est possible d'éditer n'importe quel type de dictionnaire s'il est encodé en XML.

Plusieurs projets de construction de ressources lexicales ont utilisé ou utilisent toujours cette plate-forme avec succès. C'est le cas par exemple du projet GDEF (Chalvin et al., 2006) de dictionnaire bilingue estonien-français (http://estfra.ee), du projet LexALP de terminologie multilingue sur la convention alpine (http://lexalp.eurac.edu/) ou plus récemment du projet MotÀMot sur les langues d'Asie du sud-est. Le code de cette plate-forme est disponible gratuitement en source ouverte en téléchargement depuis la forge du laboratoire LIG (http://jibiki.ligforge.imag.fr).

La plate-forme sera adaptée spécifiquement au projet DiLAF car, en sus des dictionnaires, des informations spécifiques au projet doivent être accessibles aux visiteurs :

— présentation du projet et des partenaires ;

— méthodologie générale de conversion des dictionnaires éditoriaux au format LMF (Lexical Markup Framework) (Francopoulo et al., 2006) ;

— fiches techniques concernant différents outils ou tâches à réaliser : tutoriel sur les expressions régulières, méthodologie de conversion d'un document utilisant des polices non conformes au standard Unicode vers un document conforme au standard Unicode, liste des logiciels utilisés (il s'agit uniquement de logiciels libres), méthodologie de suivi du projet ;

— présentation de chaque dictionnaire : genèse, auteurs initiaux, principes ayant régi la construction du dictionnaire, langue, alphabet, structuration des articles, etc. ;

— dictionnaire au format LMF.

Il est également envisagé de localiser la plate-forme pour chacune des langues du projet en traduisant les libellés de l'interface.

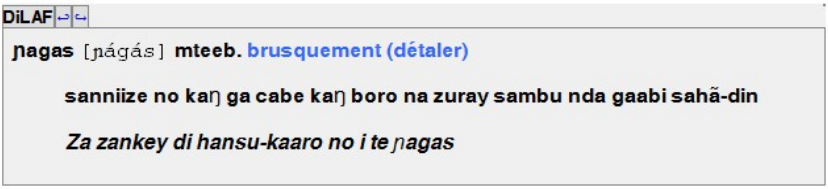

FIGURE 1 – Présentation du verbe zarma "ɲagas" sur la plate-forme jibiki

4 Travaux du premier atelier du projet DiLAF

Les participants à cet atelier sont majoritairement des linguistes ou des pédagogues, chacun travaillant sur un dictionnaire traitant de sa langue maternelle (qui est également la langue sur laquelle portent ses activités professionnelles). Les formateurs sont des enseignants-chercheurs en informatique spécialisés en traitement automatique des langues (TAL). L'objectif de ce premier atelier est de délivrer une formation à la conversion des dictionnaires tels qu'ils existent dans leur format éditorial, vers une structure XML reflétant au mieux la structure initiale des entrées tout en conservant l'ensemble des informations qui y sont exprimées. Plusieurs étapes ont été suivies pour atteindre cet objectif et garder la trace des différents traitements, chacune de ces étapes étant assortie d'un document remis aux participants.

4.1 Formation aux expressions régulières

Les participants ont été formés à l'usage des expressions régulières pendant trois jours et ont pu exercer directement leurs nouvelles connaissances par l'usage du logiciel Open Office Writer.

4.2 Conversion à Unicode

Bien que les alphabets des langues sur lesquelles nous avons travaillé soient majoritairement d'origine latine, de nouveaux caractères nécessaires pour noter des sons spécifiques à certaines langues[4] à l'aide d'un seul caractère[5] ont été adoptés par les linguistes lors d'une série de réunions[6]. La première, en septembre 1978, organisée par l'UNESCO au CELTHO (Centre d'études linguistiques et historiques par tradition orale) à Niamey crée l'« Alphabet africain de référence » fondé sur les conventions de l'IPA

[4] L'absence d'un seul signe marquant certains sons avait amené les linguistes africains à exprimer ces sons à l'aide combinaisons de lettres. Par exemple, en zarma le digraphe /ny/ note le son n palatal. C'est aussi ce qui est réalisé en français avec le son [ʃ] retranscrit /ch/.

[5] En zarma, la lettre ɲ remplace le digraphe /ny/. Ainsi, le mot autrefois écrit « nya » (mère) devient « ɲa ».

[6] Niamey (novembre 1978), Abidjan (décembre 1980), Bamako (juin 1981), Nouakchott (novembre 1981), Ouagadougou (juin 1982).

(International Phonetic Association) et de l'IAI (International African Institute). Ainsi, chacun des alphabets que nous avons précédemment présentés comprend au moins un de ces "nouveaux" caractères : ɓ ɗ ə ɛ ɣ ƙ ɲ ŋ ɔ ƴ. Des caractères composés d'un caractère latin et d'un signe diacritique ont également été créés : â ê î ô û ă ã ẽ ĩ õ ũ ḍ ḷ ṣ ṭ ẓ ǧ ǰ š ɾ.

Comme nombre de ces caractères étaient absents des dispositifs de saisie et des standards alors en usage (Enguehard, 2009), des touches de frappe de machines à écrire, des glyphes de polices d'ordinateurs ont été modifiées. Bien que la plupart de ces caractères soient depuis plusieurs années présents dans le standard Unicode (issu des travaux du comité ISO 10646 (Haralambous, 2004)), les dictionnaires dont nous disposons ont été rédigés en utilisant les anciennes polices arrangées.

Une méthodologie a été définie afin de repérer et remplacer les caractères inadéquats par les caractères définis dans le standard Unicode. Suivre cette méthodologie implique que l'ensemble des caractères repérés et leurs caractères de remplacement soient notés dans un fichier afin de pouvoir réitérer facilement cette opération si cela s'avérait nécessaire.

Ce travail est terminé et a permis de dresser la liste des caractères encore absents d'Unicode ou dont la manipulation peut poser des problèmes avec certains logiciels (voir partie 4).

4.3 Méthodologie de conversion à XML

Les fichiers électroniques des dictionnaires respectant le standard Unicode ont été convertis en fichier Open Office. Ces fichiers sont en réalité des fichiers XML compressés, les balises exprimant principalement des informations relatives à la mise en forme (usage de caractères gras ou italiques, de couleur, etc.). Il s'agit donc de passer d'un format XML dédié à l'expression de la forme vers un format XML porteur d'informations sur la structure du dictionnaire : vedette, phonétique, exemple, synonymes, etc.

Cette transformation a été partiellement ou totalement réalisée à l'aide d'expressions régulières.

5 Bilan quant à Unicode

Certains caractères des alphabets sur lesquels nous avons travaillé nécessitent d'apparaître dans le standard Unicode ou d'être mieux pris en compte par les logiciels existants.

5.1 Ordre lexicographique des digraphes

Les digraphes peuvent être facilement composés à l'aide de deux caractères mais leur usage modifie l'ordre du tri lexicographique qui conditionne la présentation des entrées du dictionnaire. Ainsi, en haoussa et en kanouri, le digraphe 'sh' est situé après la lettre 's'. Donc le verbe "sha" (boire) est situé après le mot "suya" (frite) dans le dictionnaire haoussa, et le verbe"suwuttu" (dénouer) précède le nom "shadda" (basin) en kanouri.

Ces subtilités peuvent être difficilement traitées au niveau logiciel et nécessiterait que les digraphes apparaissent en tant que signe dans le répertoire Unicode. Certains, utilisés par d'autres langues, y figurent déjà, parfois sous leur différentes casses : 'DZ' (U+01F1),

'Dz' (U+01F2), 'dz' (U+01F3) sont utilisés en slovaque ; 'NJ' (U+01CA), 'Nj' (U+01CB), 'nj' (U+01CC) en croate et pour transcrire la lettre « Њ » de l'alphabet cyrillique en serbe ; etc.

Il serait nécessaire de compléter le standard Unicode avec les digraphes des alphabets kanouri et haoussa sous leurs différentes casses.

fy	gw	gy	ky	kw	ƙy	ƙw	sh	ts
Fy	Gw	Gy	Ky	Kw	Ky	Kw	Sh	Ts
FY	GW	GY	KY	KW	KY	KW	SH	TS

TABLE 1 – Digraphes du haoussa et du kanouri absents de Unicode

5.2 Caractères avec signes diacritiques

Certains des caractères potant des signes diacritiques figurent dans une Unicode comme un unique signe, d'autres ne peuvent être obtenus que par composition.

Ainsi, les voyelles 'a', 'i', 'o' et 'u' avec tilde figurent dans Unicode sous leurs formes minuscule et majuscule1[7] tandis que le 'e' avec tilde est absent et doit être composé à l'aide du caractère 'e' ou 'E' suivi de l'accent tilde (U+303), ce qui peut provoquer des rendus différents des autres lettres avec tilde lors de l'affichage ou de l'impression (tilde situé à une hauteur différente par exemple).

La lettre j avec caron existe dans Unicode en tant que signe ǰ (U+1F0), mais sa forme majuscule doit être composée J̌ avec la lettre J et le signe caron (U+30C).

Les caractères ẽ, Ẽ et J̌ devraient être ajoutés au standard Unicode.

5.3 Editeurs de texte : fonctions changement de casse, affichage et rechercher

Les éditeurs de texte disposent généralement de la fonction changement de casse, mais ne la réalisent pas toujours de manière correcte selon les caractères. Ainsi, nous avons constaté durant nos travaux que le logiciel OpenOffice Writer (version 3.2.1) échoue dans la transformation de 'ɾ' en 'R' du bas de casse vers le haut de casse ou pour l'inverse (le caractère reste inchangé) tandis que Notepad++ (version 5.8.6) échoue dans la transformation de ǰ en J̌ du bas de casse vers le haut de casse ou pour l'inverse (le caractère reste inchangé).

Plusieurs caractères avec diacritiques peuvent être directement saisis comme un seul signe (quand celui-ci existe dans Unicode) ou être explicitement composés. Selon les logiciels, les différentes versions d'un même caractère avec diacritiques peuvent être traités de manière égale ou différente. Par exemple, le caractère 'ã', a avec tilde, peut être saisi directement comme tel (U+00E3) ou écrit comme une combinaison (U+0061 U+0303). L'affichage à l'écran avec OpenOffice Writer (version 3.2.1) est équivalent,

[7] 'ã' (U+00E3) 'ĩ '(U+0129), 'õ' (U+00F5), 'ũ' (U+O169), 'Ã' (U+00C3), 'Ĩ' (U+0128), 'Õ' (U+00D5) et 'Ũ' (U+0168).

mais la fonction rechercher appliquée à l'un de ces caractères ne permet pas de trouver l'autre ; le logiciel Notepad + + (version 5.8.6) ne permet pas d'afficher correctement les versions combinées des caractères à l'écran. La fonction rechercher ne permet pas non plus de retrouver toutes les occurrences d'un même caractère.

5.4 Caractères tifinagh

Nous complétons cet état des lieux des caractères dans Unicode par un exposé de la situation des caractères tifinagh au Niger, alphabet traditionnel des touaregs tamajaqophones.

Le tamajaq fait partie des langues berbères répartis autour du Sahara et dans le nord de l'Afrique (groupe chamito-sémitique) :

— au Maroc : tarifit au nord, tamazight au centre (Moyen Atlas), tashelḥiyt au sud et au sud-ouest (Haut et Anti-Atlas)

— en Algérie : taqbaylit au nord (Grande et Petite Kabylie), zénatya au sud (Mzab et Ourgla) chaouïa à l'est (Aurès), tahaggart des touaregs sahariens du Hoggar.

— au Mali : tamajaq de l'Adrar

— au Niger : tamajaq au nord (Aïr), au centre (vallée de l'Azawagh) et à l'ouest (le long du fleuve Niger).

Il existe également de petites communautés berbères en Mauritanie, en Tunisie ou encore en Libye (Aghali-Zakara, 1996).

Suite à une proposition marocco-franco-canadienne (Andries, 2004) des caractères tifinagh ont été introduits au sein du répertoire Unicode (Unicode, 2005), mais il apparaît qu'ils ne sont complètement adaptés à la population touarègue nigérienne utilisatrice d'alphabets tifinagh de manière traditionnelle. Au Niger, coexistent principalement deux alphabets traditionnels correspondant aux zones géographiques de l'Aïr et de l'Azawagh. Ces alphabets transcrivent 21 consonnes et la voyelle 'a' et diffèrent en ce qui concerne trois signes (Modi, 2007). De plus, ils se distinguent de l'alphabet officielle à base latinisée (voir 1.4) par l'absence de notation des consonnes emphatiques.

Valeur phonétique	Aïr	Azawagh
ɣ	⁙	⁝
q	⁝	⁘
x	⁝	⋯

TABLE 2 – Caractères divergents entre l'Aïr et l'Azawagh

De décembre 2001 à mars 2002, les caractères tifinagh ont été rénovés au Niger par un comité de linguistes spécialistes du tamajaq1[8] (Elghamis, 2003). Cet alphabet fait la

[8] Ce comité était piloté :

– à Paris par Mohamed Aghali-Zakara ;

– à Agadez par Ghoubeïd Alojaly, assisté de Emoud Salekh, Ahmed Amessalamine, Ahmed Moussa Nounou, Mohamed Adendo, Alhour Ag Analoug, Abda Annour, Aghali Mohamed Zodi, Moussa Ag Elekou ;

– à Niamey par Ramada Elghamis, avec Aghali Zennou, Ibrahim Illiasso, et Adam Amarzak.

synthèse des caractères de l'Aïr et de l'Azawagh[9] avec l'alphabet à base latine en usage pour la transcription (voir 1.4). Les linguistes ont effectué des choix là où il y avait des divergences entre les tifinaghs de l'Aïr et de l'Azawagh et fait des propositions pour la notation des voyelles ; les consonnes 'v' et 'p', utiles pour noter les emprunts, ont été ajoutées ; les signes notant les consonnes emphatiques 'ḍ', 'ḷ', 'ṣ', 'ṭ', 'ẓ' ont été construits en ajoutant un point sous le signe tifinagh correspondant (ⴹ, ⵍ̣, ⵙ̣, ⵜ̣, ⵣ̣) et les voyelles portant un signe diacritique 'â', 'ă', 'î', 'ô', 'û' ont été construites selon le même principe (ⴰ̂, ⴰ̆, ⵉ̂, ⵓ̂, ⵓ̆).

Il apparaît que l'apprentissage traditionnel de cette écriture au sein des villages facilite l'acquisition du système officiel lors de l'entrée à l'école. Par ailleurs, il existe des publications (journaux, livres) utilisant cet alphabet.

Certains caractères de cet alphabet sont absents de l'alphabet tifinagh du standard Unicode (Unicode, 2005), ou bien ont des interprétations différentes.

Caractères latins	Tifinagh APT	Unicode	Caractères latins	Tifinagh APT	Unicode
a	ⵔ	U+2D30	ŋ	ⵣ	U+2D50
ə	ⵒ	—	n	ⵏ	U+2D4F
b	ⵀ	2D40	o	ⵕ	—
c	ⵜ	—	p	ⵇ	—
d	ⵄ	U+2D39	q	ⵈ	U+2D57
e	ⵅ	—	r	ⵔ	U+2D54
f	ⵋ	U+2D3C	s	ⵙ	U+2D59
g	ⵖ	U+2D36	t	ⵜ	U+2D5C
h	ⵃ	U+2D42	u	ⵓ	—
i	ⵉ	U+2D62	v	ⵟ	—
j	ⵊ	U+2D4C	w	ⵡ	—
k	ⵅ	U+2D3E	x	ⵅⵅ	U+2D46
l	ⵍ	U+2D4D	y	ⵢ	U+2D49
m	ⵎ	U+2D4E	z	ⵣ	U+2D63

TABLE 3 – Caractères tifinagh APT (sans signe diacritique) et Unicode

Ce recensement fait apparaître l'absence de huit caractères dans le standard Unicode.

6 Futurs travaux

Les futurs travaux du projet DiLAF porteront dans un premier temps sur la correction des erreurs relevées dans les dictionnaires, et l'ajout d'entrées manquantes relatives aux mots désignés par les liens de synonymie, d'antonymie, etc.

la seconde étape consiste en un enrichissement des dictionnaires afin d'être en mesure de

[9] Le signes 'j' en est absent.

calculer toutes les formes fléchies des noms et adjectifs et toutes les conjugaisons des verbes.

Dans la mesure du possible une troisième étape de traduction des exemples et définitions vers une ou plusieurs autres langues sera définie afin de constituer des corpus plurilingues.

7 Conclusion

Le projet DiLAF établit une méthodologie de conversion de dictionnaire éditoriaux vers des formats XML. Il s'agit de créer et rendre disponibles de nouvelles ressources aux chercheurs en TAL, d'une part et de d'équiper les langues africaines de ressources numériques nouvelles et indispensable à leur promotion, d'autre part.

La publication de ces ressources sur Internet permettra aux locuteurs de ces langues de disposer, souvent pour la première fois, d'informations linguistiquement fiables quant à l'orthographe, au lexique ou vocabulaire et à l'usage des mots de leur langue.

La tenue de ce premier atelier a permis de rassembler une dizaine de linguistes de trois pays ainsi que deux informaticiens. Les travaux menés ensemble ont fait émerger la richesse de la collaboration entre disciplines complémentaires et entre pays voisins. Les transferts de connaissance ont été riches, tant en ce qui concerne les outils techniques que sur des sujets de fond en linguistique. Les formations communes, les réalisations de chacun et les discussions ont fait émerger une synergie d'action entre les pays concernés.

8 Remerciements

Nous remercions spécialement M. Moukeïla Sanda, à l'initiative de ce projet, Mme Rabi Bozari, directrice de l'Institut National de Documentation, de Recherche et d'Animation Pédagogiques, Mme Rakiatou Rabé, M. Maï Moussa Maï et Mahamou Raji Adamou, linguistes, sans qui ce projet ne pourrait être mené à bien.

Le projet DiLAF[10] est financé par le Fonds Francophone des Inforoutes de l'Organisation Internationale de la Francophonie.

9 Références

Aghali-Zakara, M. (1996).Eléments de morpho-syntaxe touarègue. CRB / GETIC.

République du Niger. (1999). Alphabet haoussa, arrêté 212-99.

République du Niger. (1999). Alphabet kanouri, arrêté 213-99.

République du Niger. (1999). Alphabet tamajaq, arrêté 214-99.

République du Niger. (1999). Alphabet zarma, arrêté 215-99.

Andries, P. (2004). Proposition d'ajout de l'écriture tifinaghe. *Organisation internationale de normalisation.* Jeu universel des caractères codés sur octets (JUC). ISO/IEC JTC 1/SC 2 WG 2 N2739.

[10] http://www.inforoutes.francophonie.org/projets/projet.cfm?der_id=262

Berment, V. Méthodes pour informatiser des langues et des groupes de langues peu dotées. Ph.D. thesis, Université Joseph Fourier, 2004.

Chalvin, A. et Mangeot, M. (2006). Méthodes et outils pour la lexicographie bilingue en ligne : le cas du Grand Dictionnaire Estonien-Français. *Actes d'EURALEX 2006*, Turin, Italie, 6-9 septembre 2006, 6 pages

Elghamis, R. (2003). Guide de lecture et d'écriture en tifinagh vocalisées. *APT*, Agadez, Niger, janvier.

Eluerd, R. (2000). La Lexicologie. Paris, PUF, Que sais-je ?

Enguehard, C. (2009). Les langues d'Afrique de l'Ouest : de l'imprimante au traitement automatique des langues*, Sciences et Techniques du Langage*, 6, pages 29-50, p.29-50,. (ISSN 0850-3923).

Francopoulo F., George M., Calzolari N., Monachini M., Bel N., Pet M. et Soria C. (2006). Lexical Markup Framework (LMF). *LREC 2006 (International Conference on Language Resources and Evaluation)*, Genoa.

Haralambous, Y. (2004). Fontes & codages, O'Reilly France.

Mangeot, M. (2001). Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue. Thèse de nouveau doctorat, Spécialité Informatique, Université Joseph Fourier Grenoble I, 280 pages, jeudi 27 septembre.

Mangeot, M., Sérasset, G. et Lafourcade, M. (2003). Construction collaborative de données lexicales multilingues, le projet Papillon. *Revue TAL, édition spéciale, Les dictionnaires électroniques : pour les personnes, les machines ou pour les deux ? (Electronic dictionaries: for humans, machines or both?)* Ed. Michael Zock & John Carroll, Vol. 44:2/2003, pages 151-176.

Mangeot, M., et Thevenin, D. (2004). Online Generic Editing of Heterogeneous Dictionary Entries in Papillon Project. *Proc. of COLING 2004*, ISSCO, Université de Genève, Switzerland, 23-27 August 2004, vol 2/2, pages 1029-1035.

Mangeot, M., et Chalvin, A. (2006). Dictionary Building with the Jibiki Platform: the GDEF case. *Proc. of LREC 2006*, Genoa, Italy, 23-25 May 2006, pages 1666-1669.

Matoré, G. (1973). La Méthode en lexicologie. Paris, Didier.

Modi, I. (2007). Les caractères tifinagh dans Unicode. *Actes du colloque international "le libyco-berbère ou le tifinagh : de l'authenticité à l'usage pratique"*, pages 241-254, ed. Haut Commissariat à l'amazighité (HCA), pages 21-22, mars, Alger.

Mortureux, M.-F. (1997). La lexicologie entre langue et discours. Paris, SEDES.

Streiter, O. Scannell, K. P. et Stuflesser, M. (2006). Implementing NLP Projects for Non-Central Languages: Instructions for Funding Bodies, Strategies for Developers. *Machine Translation*, vol. 20 n°3, March.

Unicode (2005). The Unicode Standard 4.1, Tifinagh, range 2D30-2D7F.