

Études et Documents Berbères, 22, 2004 : pp. 175-173

L'ÉCRITURE AMAZIGHE TIFINAGHE ET UNICODE

par

Lahbib Zenkouar

Le dossier Unicode a démarré depuis la confrontation des informaticiens du Centre des études informatiques et des systèmes d'information et de communication de l'Institut royal de la culture amazighe (IRCAM) du Maroc, avec l'écriture des programmes pour la réalisation du clavier et des polices de caractères de l'alphabet amazigh, version IRCAM, proposé par le Centre d'aménagement linguistique. Les études menées au centre des études informatiques sur cette problématique ont conclu à l'évidente nécessité d'avoir un codage informatique pour l'alphabet amazighe.

Parmi la panoplie des codages, celui qui représente la meilleure possibilité est, sans aucun doute, le standard de codage Unicode et ceci pour de multiples raisons. C'est un codage ouvert par opposition aux autres nombreux systèmes de codage. En plus, Unicode a l'intéressant avantage d'englober tous les autres systèmes d'encodage. C'est un codage ouvert en ce sens que d'éventuelles révisions peuvent être opérées sur celui-ci ; ce qui est très rare, sinon impossible, dans les normes de codage antérieures.

I. LE SYSTÈME DE CODAGE UNICODE ET LES SYSTÈMES ANTÉCÉDENTS

Les systèmes qui précèdent à Unicode dérivent et sont des variantes du célèbre code ASCII. Celui-ci, spécifiquement américain, a été le standard de codage des caractères depuis plus de 30 ans. Conçu spécifiquement pour la langue anglaise, il a posé un certain nombre de problèmes de représentation de caractères en informatique aux autres langues. Mais, il faut se rappeler qu'au début de l'ère informatique, on pensait plutôt à la réussite de l'informatique en tant qu'outil technologique plutôt que celui d'un moyen de communication multilingue intégrant au mieux les droits fondamentaux des langues et écritures [1], [2].

Au début des années 1990, deux initiatives de remplacement du code ASCII par un nouveau code universel de caractères ont été lancées, l'une par l'ISO intitulée UCS (Universal Character Set) et l'autre par un consortium de fabricants, Unicode. Les deux entités se sont entendues pour mener cette entreprise en commun depuis 1993. Pour être plus précis, la définition d'Unicode est calquée sur la norme ISO/CEI 10646, UCS-2, c'est-à-dire la norme ISO/IEC 10646-1:1993

Le standard Unicode se définit comme un système de codage mettant en œuvre un mécanisme cohérent et universel de codage de caractères permettant ainsi aux textes multilingues de coexister. Ce système facilite l'échange de données textuelles à l'échelle planétaire et permet d'engendrer et de créer une informatique à l'échelle internationale. C'est un point positif qui montre, si besoin est, l'ascension encore exponentielle de l'informatique.

Grâce à ce nouveau standard Unicode, l'industrie informatique assurera la stabilité des caractères et permettra de simplifier le développement des logiciels et leur internationalisations. En outre ce système de codage universel réduit fortement les coûts du développement. Il faut quand même signaler que bien qu'il ait été normalisé, le code UCS/Unicode est seulement depuis peu réellement mis en œuvre [3], [4].

Par ailleurs, le standard Unicode est devenu le codage supportant les langages de mise en page électronique tels que HTML et XML. Il constitue de ce fait le codage utilisé pour l'Internet et les nouveaux protocoles relatifs aux technologies de l'information et de la communication tels que normalisés et standardisés par le consortium W3C [3], [4], [5].

Dans cet article où je m'attacherai à décrire les propriétés et les formats de stockage et de transfert utilisés par le système de codage Unicode pour introduire ensuite le codage de l'alphabet amazigh, je commencerai tout d'abord par décrire le modèle de codage des caractères, ensuite je décrirai leurs propriétés et finalement les formes normalisées des caractères dans ce système de codage pour exposer la place de l'amazigh dans ce système de codage.

1. Le modèle de codage des caractères

Le modèle de codage s'appuie, autant que faire se peut, sur des définitions aussi précises que possible des lettres des écritures de notre monde ; en ce sens que chaque notion qui se révèle imprécise n'est plus utilisée malgré son maintien dans les tableaux de codage afin d'éviter toute ambiguïté dans les formes de codage qui s'ensuivront. Les possibilités d'amendement aux différentes propositions soumises aux instances de l'ISO/CEI permettent de pallier aux différents problèmes pour que chaque langue trouve son compte sur cet outil.

Ces organismes de normalisation ont instauré une procédure d'examen des propositions à plusieurs niveaux. Cette manière de procéder rend un grand service à la mise au point définitive d'un projet d'amendement de sorte que la présence d'éventuelles erreurs reste relativement peu fréquente.

Pour ce système, nous définissons d'abord les notions suivantes [4], [5], [6], [7] :

Le répertoire. C'est un ensemble de caractères et non de glyphes. Il peut ne pas être ordonné. En plus, le système de codage Unicode maintient ce répertoire ouvert pour tout nouvel examen de ce même répertoire.

Le jeu de caractères codés. Ce répertoire devient un jeu de caractères codés dès le moment où on établit une correspondance entre l'ensemble des caractères abstraits du répertoire et un ensemble d'entiers positifs pour définir le répertoire de manière numérique. Bien entendu, ce dernier ensemble peut ne pas être contigu. En définitive et de manière générale, nous dirons qu'un caractère abstrait est codé dans un jeu de caractères donné si un numéro est associé à ce caractère.

2. La forme en mémoire

C'est la forme des caractères stockés en mémoire. Cette forme naturelle en mémoire s'appuie sur les unités de stockage. L'unité de stockage est un nombre binaire d'une certaine largeur ou nombre de bits (octet, seize...) qui sert d'unité de base à l'expression des numéros de caractère en mémoire de l'ordinateur.

Le système Unicode utilise la norme à 16 bits de référence dans l'industrie de l'internationalisation ; il s'agit d'un code de longueur fixe pouvant représenter toutes les langues modernes. La forme UCS-2 est la forme 16 bits de l'ISO/CEI 10646 qui est équivalente à celle du consortium Unicode définie sur 16 bits.

Les deux formes d'encodage utilisé dans ce système qui se veut multilingue sont :

UCS-n : la forme à n bits de l'ISO/IEC 10646 qui traduit la forme naturelle de représentation binaire de l'information en mémoire.

UTF-n : UCS Transformation Format à n bits, un code d'échange et de transfert défini pour rendre pratique et opérationnel la forme de l'ISO 10646 et qui permet d'exploiter la forme octale utilisée par la plupart des systèmes actuels.

Par conséquent, les données Unicode peuvent être codées sous deux formes : une forme naturelle de 16 bits et une forme de 8 bits (UTF-8) conçue pour faciliter une utilisation sur les systèmes ASCII préexistants comme nous venons de le signaler. En effet, l'intérêt pratique de l'UTF-8 est sa capacité à préserver les standards existants. Cette forme d'encodage qui s'impose de plus en plus, est utilisée par Java, Netscape, Oracle, etc.

Ces méthodes d'encodage conservent la taille d'encodage pour chaque caractère, ce qui est fondamentalement différent des anciennes méthodes, qui utilisaient, par exemple des balises de contrôle pour passer d'un code de caractères à un autre et qui constituaient des variantes de l'ASCII. En outre, ceci revêt une grande importance lorsque la taille des documents devient importante. Un texte en latin encodé en Unicode est deux fois plus grand que le même texte encodé en ASCII. En revanche, l'encodage en Unicode d'un texte asiatique produit sensiblement la même quantité d'octets [3].

En définitive, nous pouvons résumer les formes d'encodage UCS, par le fait qu'elles permettent de coder tous les caractères de toutes les langues. UCS se compose essentiellement de tables de codes et de noms de caractères, tandis que l'équivalent de l'ISO, en l'occurrence Unicode y ajoute des précisions et des recommandations d'ordre typographique.

Selon les mises en oeuvre, chaque caractère est exprimé au format UCS-2 (deux octets); c'est le plan multilingue de base (BMP) ou au format UCS-4 (quatre octets, code complet) réunissant le plan multilingue de base et les plans complémentaires. Lorsqu'une mise en oeuvre se heurte à une incompatibilité, la méthode de transformation UTF-8, vue précédemment, est disponible.

3. Les propriétés de caractères

Il s'agit d'un certain nombre d'informations sur les caractères destinées à mieux servir Unicode. On distingue les propriétés normatives qui sont la catégorie auquel appartient le caractère et les propriétés informatives destinées elles à renseigner le développeur sur l'importance et les subtilités des caractères de la langue considérée sans que celui-ci soit obligatoirement un locuteur de cette langue.

Ces propriétés servent à l'automatisation du traitement informatique de la langue. Les propriétés des caractères sont utilisées dans les algorithmes de rendu des caractères, de tri à base de classement des chaînes de caractères, de découpe en ligne et, en général, de tout ce qui peut identifier un ensemble de caractères tels que les caractères liants ou antiliants avec ou sans chasse etc.

Les propriétés Unicode concernent toutes les propriétés répertoriées par le système de codage Unicode. En général, nous avons la catégorie générale, la classe bidirectionnelle (direction d'écriture de la lettre), la décomposition, les classes combinatoires canoniques, la gestion des liaisons.

La liste officielle des noms de caractères selon leur position dans le codage se trouve dans un fichier texte à l'adresse: <http://www.Unicode.org/public/UNIDATA/UnicodeData.txt> [6]. Ce fichier texte contient un grand nombre d'informations formatées de manière à être traitées par les programmes informatiques. Une version française de ce fichier se trouve dans le site Hapax canadien. Il y'a au total quinze champs pour définir un caractère [6].

a) Catégorie générale

C'est la propriété la plus importante d'un caractère, celle qui va déterminer son comportement dans la plupart des traitements linguistiques et typographiques.

Nous citerons pour exemple les catégories suivantes [2], [4], [5] :

Les lettres (majuscule, minuscule, titrage, modificatrice, autre type), les signes diacritiques ou marques (sans chasse, combinatoire à chasse non nulle, englobante), les nombres (décimal, alphabétique, autre), les ponctuations (connecteur, tiret, ouverture, fermeture, guillemet ouvrant et fermant, autre), les symboles (mathématique, monétaire, modificateur, autre), les séparateurs (espace, de lignes), etc.

Unicode fait l'effort de classer les langues du monde autant que faire se peut. La complexité des langues et leurs ambiguïtés a obligé les rédacteurs de normes à inventer d'autres rubriques pour classer les caractères particuliers très souvent des langues non latines. On trouve également dans cet effort louable de classification les deux caractéristiques suivantes qui peuvent intéresser l'écriture amazighe de l'Afrique du Nord, ce sont les extenseurs dont le rôle est d'étendre ou de réitérer le caractère qui les précède, ou relatifs à la gestion des liaisons par les liants sans chasse ou avec chasse.

b) Décomposition

Unicode code parfois un même caractère sous deux formes différentes mais dont l'apparence est identique. Il s'agit de la forme précomposée, essentiellement pour des raisons historiques, avec un seul point de code et une autre forme décomposée, constituée par la combinaison du caractère de base et du caractère diacritique. Cette forme, considérée par les instances de normalisation comme plus productive, est vivement encouragée par le système Unicode par rapport à la forme précomposée consommatrice d'emplacements mémoire. Exemple : U+00C5 (Å) // U+0041 (A) U+030A (°). Cette forme de codage bien qu'économique peut cependant créer des problèmes linguistiques pour certaines langues et poser des problèmes techniques dans le traitement informatique et même linguistique de ces caractères.

c) Équivalence canonique

L'équivalence canonique se rencontre lorsque des caractères sont considérés comme identiques et ne diffèrent pas au niveau visuel. Nous donnons comme exemple le caractère É (U+00C9) qui est une variante canonique de E (U+0045) + ' (U+0301)

De manière formelle, on dit que deux suites de caractères sont des équivalents canoniques si leurs décompositions canoniques, complètes et récursives respectives, sont identiques.

d) Classes combinatoires

On appelle caractères combinatoires ou combinants des éléments qui s'associent à un caractère de base. Il s'agit essentiellement des accents et autres signes diacritiques qui se combinent à une lettre pour en constituer une autre. Les diacritiques se retrouvent dans de nombreuses langues notamment l'arabe, l'hébreu, le latin, etc.

Le caractère de base se retrouve avec différentes positions réservées aux diacritiques qui constituent les classes. Ces classes sont utilisées par l'algorithme de mise en ordre canonique défini par le standard Unicode. Ils sont identifiés par des valeurs numériques décimales : 0 : Avec chasse, fendues, englobantes, antéposées ; 1 : Couvrantes et intérieures ; 7 : Nouktas ; 208 : Antéposées et jointes à gauche ; 218 : Souscrites à gauche ; 224 : Adscrites à gauche (d'un seul car de base) ; 230 : Suscrites ; etc.

Par ailleurs, nous pouvons rencontrer les cas suivants :

- Dans le code UCS, tous les signes diacritiques ont un code qui leur est propre et ils peuvent être combinés avec n'importe quel autre caractère.
- Ces signes sont dits « à chasse nulle ». Ainsi, un « À » peut aussi bien être obtenu par le recours à son code spécifique (U00C0) que par combinaison d'un A majuscule (U0041) avec le code de l'accent grave à chasse nulle (U0300).

e) Propriétés de composition

Cette partie traite de la nature des caractères : caractères simples ou composites, caractères normaux ou caractères de compatibilité, symboles combinatoires espaçant ou non espaçant. Il fait le point sur le processus de composition et de décomposition et sur les questions de normalisation et d'ordonnancement.

f) Décomposition canonique

Certains caractères ont la propriété d'être décomposables en une séquence de caractères élémentaires. Ce sont les caractères composites, composés ou précomposés. C'est le cas des lettres accentuées en français. Le *e* est considéré comme une lettre de base et les accents comme des marques additives qui reçoivent le nom de marques combinatoires.

Dans la plupart des cas les marques ne comportent pas d'espace, mais il existe des marques combinatoires avec espace.

Il résulte de toutes ces considérations que les applications supportant Unicode doivent être capables de réaliser dans certaines circonstances des décompositions de caractères composites, ou inversement de recomposer des séquences combinatoires, pour trouver un composite correspondant. Ce sont les opérations de composition et de décomposition. La décomposition se fait

en utilisant des décompositions fournies par le fichier de propriétés des caractères Unicode. Il faut effectuer cette opération de manière répétitive jusqu'à ce que la séquence obtenue ne comporte plus que des éléments indécomposables. La forme à laquelle on aboutit s'appelle *décomposition canonique* ou *composite canonique*.

Nous pouvons finalement accorder les caractéristiques suivantes aux caractères canoniques décomposés : ils sont réversibles et ne subissent pas de perte d'information.

Cette forme est utilisée dans l'échange normalisé de textes. En effet, cette forme permet d'effectuer une comparaison binaire tout en conservant une équivalence canonique avec le texte non normalisé d'origine

g) Décomposition de compatibilité

Les choses sont moins rigoureuses lorsqu'on aborde la question des caractères de compatibilité. Ce sont des caractères qui ne devraient pas figurer dans le catalogue Unicode mais qui y ont été acceptés afin de préserver une correspondance avec les encodages préexistants pour, par exemple, garantir des conversions parfaites d'un standard à un autre. Les décompositions de compatibilité se caractérisent par une perte d'information visuelle (pas exactement la même apparence) mais permettent d'effectuer une comparaison binaire tout en conservant une équivalence de compatibilité avec le texte non normalisé d'origine. Ce fait s'avère utile car il permet d'éliminer des différences qui ne sont pas toujours pertinentes

h) Normalisation

Étant donné la multiplicité des représentations possibles pour un même caractère, certaines tâches relatives au traitement de texte se trouvent compliquées. Les opérations de recherche d'une chaîne à l'intérieur d'un texte ou les opérations de comparaison et par voie de conséquence de tri en sont des exemples. On ne peut se contenter d'effectuer des opérations sur les octets en mémoire puisqu'un même caractère pourrait être représenté par un nombre variable d'octets selon qu'il est sous forme composite ou sous forme partiellement ou totalement décomposée.

Il est donc indispensable de définir une forme canonique sous laquelle mettre les caractères afin d'éliminer les ambivalences. Une telle forme doit garantir une représentation unique pour chaque caractère.

Le processus de transformation qui met chaque caractère d'une chaîne sous une telle forme canonique est appelé *normalisation*. La chaîne transformée est dite sous forme normalisée. Le standard Unicode définit quatre types de normalisation.

En conclusion, nous pouvons résumer ces notions en disant qu'afin de garantir une représentation unique de ce qui est considéré comme équivalent canonique ou de compatibilité, Unicode définit parfois plusieurs codes qui correspondent :

- à des entités qui peuvent être considérées comme identiques (variantes canoniques) ;
- ou qui ne sont que des variantes visuelles d'un même caractère (variantes de compatibilité).

	<i>Sans composition canonique</i>	<i>Suivie d'une composition canonique</i>
<i>Décomposition Canonique</i>	<i>D</i>	<i>C</i>
<i>Décomposition de compatibilité (K)</i>	<i>KD</i>	<i>KC</i>

II. LE VOLET INFORMATIQUE ET TECHNIQUE DE L'AJOUT DE TIFINAGHE À UNICODE

Historique de la proposition

Suite à une réunion avec M. Mohamed Chafiq, académicien du Royaume du Maroc et premier Recteur de l'IRCAM, sur la nécessité et la problématique du codage des tfinaghes pour l'inscription de la langue amazighe dans l'ère des nouvelles technologies, celui-ci a compris l'importance de ce dossier et a encouragé le Centre des études informatiques et des systèmes d'informations et de communication (CEISIC) à entamer et à réussir ce dossier de première importance.

C'est ainsi que des contacts ont été pris avec les experts de l'ISO et du consortium Unicode pour confirmer une éventuelle proposition faite par un organisme au sujet de l'alphabet amazighe comme le laissait croire certaines rumeurs sur une éventuelle proposition déposée par IBM Égypte. En outre, certains chercheurs confondaient le dossier du codage électronique de la graphie amazighe et la reconnaissance de cette graphie comme système d'écriture de la langue amazighe. Ce qui est une autre norme nationale de la codification de l'écriture de la langue amazighe déposée auprès de l'organisme de normalisation marocain principalement par le centre d'aménagement linguistique. Cette démarche est venue, suite à une réunion de travail où nous avons pris la décision de déposer cette norme nationale concernant l'écriture de la langue amazighe[10], pour fortifier et assurer la réussite de la proposition de codage, qui est l'objet de cet article.

Après les recherches auprès de l'ISO, les experts contactés n'étaient au courant d'aucune proposition de codage provenant d'un quelconque orga-

nisme ou d'un pays. Le directeur du Centre informatique de l'IRCAM a pris contact avec plusieurs chercheurs dans ce domaine de la normalisation informatique tant nationaux qu'étrangers. C'est ainsi que les canadiens Alain La Bonté, Patrick Andries et François Yergeau nous ont contacté pour s'enquérir de nos objectifs. Ces experts, qui s'intéressent à la codification des langues, étaient des membres soit du consortium Unicode soit de l'ISO ou des deux à la fois [8]. Ces scientifiques avaient, par ailleurs à titre volontaire, déjà travaillé sur la problématique du codage des écritures pour noter la langue amazighe et étaient au fait de son importance comme une grande et ancienne écriture de toute l'Afrique du Nord.

En accord avec le Centre d'aménagement linguistique, qui a le mérite d'avoir confectionné, à partir de notre écriture ancestrale, un ensemble restreint de signes pour la standardisation des parlers amazighes du Maroc, nous avons accueilli M. Patrick Andries qui était disponible pour venir nous rendre visite à Rabat. Nous avons alors tenu plusieurs réunions de travail avec ce chercheur avec qui nous avons clarifié plusieurs points de vue et mis au point les stratégies de travail pour réussir cet important dossier. Certains détails informatiques et techniques devaient attendre une concertation plus large. Le travail scientifique avait démarré depuis ce temps : M. Andries s'est chargé de la rédaction des résultats de nos exigences et recommandations. Cette rédaction s'appuierait sur la graphie tifinaghe proposée par le Centre d'aménagement linguistique et utilisée depuis au Maroc et des autres variantes qui prendraient en charge quelques particularités régionales au Maroc et en Afrique du Nord de manière générale. Cette proposition est écrite également de manière à refléter les caractéristiques informatiques et techniques des points de code [8], [9].

Outre le fait que le Centre informatique est à la base de la proposition de codage électronique et informatique (ISO/CEI 10646) de l'alphabet amazighe et son principal maître d'œuvre [13], il a contribué scientifiquement au montage de la proposition et a été un élément incontournable au cours des discussions avec l'ISO et les prises de décisions qu'il avait l'entière charge de prendre et d'en assumer les conséquences dans le cadre des ses prérogatives de centre informatique au sein de l'IRCAM. Le staff technique de ce centre comporte des chercheurs et ingénieurs en systèmes de communication formés dans cette science de codage. Les membres du conseil d'administration de l'IRCAM avaient bien raison de proposer d'élever au rang de centre de recherche le département technique et informatique qui était voué au départ à une gestion du matériel technique et proposé comme tel dans la première version de travail du règlement intérieur de l'IRCAM. Le Professeur Chafiq a aussitôt compris l'injonction des membres du conseil d'administration et a défendu ce point de vue avec lequel il était entièrement d'accord. C'est à ce changement qu'on doit l'existence d'un centre s'occupant de l'informatique et des technologies de l'information au sein de l'IRCAM.

Caractères et glyphes

Bien que la philosophie et les concepts à la base d'Unicode ont été mises au point dans ce système de codage Iso/Unicode, il n'en demeure pas moins que depuis l'apparition de ce standard, capable de décrire toutes les écritures du monde, une définition plus précise du caractère s'est développée au fur et à mesure des cas traités à l'occasion des propositions de normes et de leurs nombreux amendements. Elle a pu séparer entre le caractère et sa représentation par un dessin ou le glyphe. Cette définition est principalement le résultat des études et analyses des écritures orientales qui changeaient de formes et possédaient plusieurs glyphes pour un même objet de leur alphabet.

L'écriture informatique utilise les codages et les fontes (ensemble de glyphes). Ceux-ci tout en étant liés, répondent à des besoins différents. Le codage est une notion informatique de l'écriture qui a pour but de caractériser l'information dans ses composants élémentaires.

En fin de compte, le codage d'un alphabet est une table de caractères avec un ensemble de points de code manipulable par tout outil informatique ou support de stockage. Le caractère défini par son numéro ou point de code est un élément informatique qui représente une notion abstraite au sens de l'alphabet. Un glyphe est une image ou un dessin donnant forme à ce caractère informatique abstrait.

En entrant dans cet espace de l'écriture informatique, l'utilisateur participe, sans en être conscient, à un va et vient incessant entre caractères et glyphes. Quand il appuie sur une touche, le code informatique du caractère est transmis du clavier et le glyphe correspondant s'affiche à l'écran. En faisant de la recherche morphologique dans un texte, l'ordinateur affiche des glyphes à l'écran [3].

L'intervention des linguistes consistait à spécifier les glyphes adoptés pour la graphie tiffinaghe et à référencer leurs choix. L'ensemble des autres caractères non adoptés par l'IRCAM étaient discutés et s'appuyaient sur des références amazighes et d'illustres chercheurs de toute l'Afrique du Nord.

Au cours d'une réunion de travail qui a réuni les chercheurs du cal, du ceisic et l'expert canadien, en la personne de M. Andries, nous avons convenu d'étendre la graphie proposée par le centre d'aménagement linguistique à un alphabet qui prendrait en charge certaines lettres correspondantes à des particularités phonétiques régionales et quelques lettres latines non couvertes par notre alphabet. L'intérêt est de satisfaire les chercheurs qui envisageraient d'étudier la langue amazighe au Maroc dans toutes ses expressions [10].

C'est ainsi que la proposition comprend quatre sous-ensembles de caractères tiffinaghes : le jeu de base et le jeu étendu de l'IRCAM qui sont inspiré complètement de notre écriture ancestrale, des lettres néo-tiffinaghes en usage et des lettres touarègues modernes dont l'usage est attesté.

La collaboration avec les experts canadiens avait pour objectif d'aboutir à une bonne rédaction de la proposition de la norme et à son introduction auprès de l'ISO vu leur expérience des rouages de ces institutions internationales.

Le codage de l'écriture amazighe

Nous avons soulevé plusieurs aspects relatifs au codage de l'alphabet amazighe. Le premier est le fait que la langue amazighe est une langue vivante malgré une conjoncture très difficile qui a duré plus d'un demi-siècle. De ce fait, elle méritait et devait être codée dans le plan multilingue de base ; c'est-à-dire sur 16 bits en UCS2 (ou son équivalent UTF-16) et sa variante UTF-8 qui permettra à cette langue de traverser tous les interfaces préexistants à la définition de la norme 10646 et de lui assurer une utilisation plus simple et surtout pratique.

Nous avons insisté sur le fait que les caractères amazighs devaient être codés de manière consécutive, ce détail très utile permettra de simplifier des traitements automatiques potentiels, en particulier la reconnaissance automatique de la langue amazighe. Nous craignons que l'encombrement au niveau du plan multilingue de base ne donne un codage éparpillé, ce qui compliquerait les opérations de traitement de l'information sur la langue. C'est ainsi que la plage de point de code réservés à l'amazighe occupe l'espace hexadécimal 2D30 à 2D7F.

Les phonèmes à diacrité ont été aménagés de sorte que la diacrité soit indépendante et considérée comme un caractère indépendant. Malgré sa petite taille transposée pratiquement en exposant, le signe diacritique constitue un caractère à part entière. Ce que nous voulons obtenir est le fait que les lettres à diacrité ne soient pas des éléments combinants au sens d'Unicode. Ils n'auront pas besoin d'une décomposition d'équivalence. La seule décomposition possible est canonique et possède des valeurs cibles biunivoques. De ce fait, tout texte écrit en alphabet amazighe marocain sera naturellement normalisé par décomposition canonique et peut subir tous les traitements de l'information, comme par exemple le traitement de texte [3], [4]. De toutes les manipulations possibles et permises par le langage de description et de transfert de données et en particulier XML. De ce fait, la proposition telle que rédigée ne mentionne pas de signes diacritiques [10].

Nous avons agi de sorte que cet alphabet s'adapte parfaitement aux recommandations de l'organisme régulant et organisant les technologies du *Web*, le *world wide web* consortium : W3C. Nous avons eu plusieurs échanges sur ce sujet avec le professeur Najib Tounsi, responsable du bureau W3C marocain, ainsi qu'avec M. François Yergeau, un spécialiste canadien confirmé et M. Rishard Ishida, un officiel britannique à l'échelle internationale de cet

organisme [3], [4], [5]. En fait, ce point sera pris en compte dans la future table commune de tri tirée principalement de la norme ISO 10646.

Le fait que la diacrité soit codée par le nombre hexadécimal 2D6F, c'est-à-dire neuf emplacements plus loin du dernier caractère amazighe (2D65) est choisi pour faciliter les opérations de tri. C'est un point très important que nous avons soulevé dans le codage lors de nos discussions avec les rédacteurs et experts de ces instances. Ce point permettra de classer la lettre diacritique au second plan après la même lettre sans signe diacritique [6]. Ceci aura pour effet de faciliter les opérations de tri de cet alphabet.

En outre, et c'est un avantage de plus dans le dossier Unicode du Tifinaghe, nous n'avons aucune forme de ligature susceptible de créer une ambiguïté de normalisation. Il existe cependant des formes de ligatures en berbère, mais nous avons préféré ne pas les utiliser pour avoir un alphabet amazighe aussi clair et précis que possible.

Un autre point positif en ce sens qu'il ne complique pas l'opération de codage, est l'absence de glyphes marquant le début de phrases; en d'autres termes par ce que les écritures latines dénomment *majuscules*. L'écriture amazighe prévoit de marquer la fin de la phrase par un point. Il est alors redondant de marquer le début de la phrase suivante par la majuscule. L'influence du côté informatique est de taille, il faut une couche soft en dessus du simple codage pour spécifier ce glyphe car nous ne pouvons pas prétendre à l'historicité des normes de codage antérieures à Unicode pour imposer des points de code pour les majuscules comme en ont profité les occidentaux. En résumé, la langue amazighe ne peut que s'en porter mieux car ces glyphes, représentant les majuscules, ne manqueraient pas de surcharger les tests pour les traitements relatifs aux textes amazighes dans toutes les applications informatiques futures : tri, traduction, vérification, etc.

Nous n'avons pas réservé d'emplacements pour d'éventuelles majuscules à inventer. Du coup le questionnement de certains chercheurs sur la nécessité d'introduire les majuscules n'a plus droit au chapitre. Ceci n'empêche pas un développeur de donner un format de police avec un corps plus important pour marquer le début d'une phrase ce qui ne nécessite pas d'associer les codes points. Il est n'est pas impossible d'utiliser des glyphes pour les majuscules sans avoir besoin d'un code.

Par la suite, les linguistes du Centre d'aménagement linguistique, nous ont soutenu dans notre effort de simplification et de normalisation de la graphie amazighe telle que proposée. La palatalisation était tout simplement supprimée des extensions proposées; ne restait à prendre en charge que les labiovélares avec le même signe diacritique sans en constituer véritablement un au niveau du codage Unicode comme nous l'avons expliqué précédemment.

Dans les mêmes discussions qui ont duré pratiquement une année, les

formes possibles du clavier amazighe, suite aux propositions techniques sur le codage, ont été âprement discuté sur tous les points de vue avec les experts de ces instances. Nous voulions la forme la plus simple de notre clavier. La diacrité codée indépendamment arrangeait bien le problème. Mais nous voulions également outrepasser la touche morte ou de *switch* du clavier qui permettait d'avoir les sept autres caractères dans la version locale actuelle s'appuyant sur le code ACSII du clavier amazighe version IRCAM. De 33 lettres, nous sommes passés à 32 caractères. D'ores et déjà, nous pouvons annoncer que le prochain clavier sera différent de l'actuel et plus simple [11].






Grâce à la collaboration internationale et à la correspondance électronique, une large consultation a eu lieu autour de la proposition centrale initiée par l'IRCAM. Elle a rallié la France grâce à ses nombreuses institutions étudiant et délivrant des diplômes en langue amazighe, le Canada par la présence d'experts canadiens et les nombreux chercheurs de l'Afrique du Nord et du Sahel. La liste bibliographique de la proposition est loin d'être exhaustive et ne reflète pas toutes les interventions de toutes ces contrées en Afrique et à l'étranger [12].

La proposition rédigée par Patrick Andries, suite à nos discussions, mentionnait le Maroc comme principal soumissionnaire suivi du Canada et de la France. Elle était soutenue par de nombreux autres pays [8], [9]. Les références bibliographiques sont éloquents sur son caractère pan-africain. Pour notre part, nous avions pour seul souci la réussite de cette importante et stratégique opération qui inscrirait définitivement la langue amazighe dans le troisième millénaire qui est tout simplement l'ère de l'informatique. C'est à ce titre que nous nous n'avons pas cessé de rappeler la commission scientifique de l'IRCAM sur l'importance à accorder à ce volet et pesé de tout notre poids pour sa réussite. Pour l'histoire, notre correspondance avec les experts sur les aspects de la normalisation dépasse facilement les trois cents emails pour le seul dossier du codage.

Partis d'une simple proposition de l'IRCAM initié par le centre informatique sur un substrat alphabétique proposé par le Centre d'aménagement linguistique, nous avons abouti à l'instruction d'un dossier international. C'est à notre sens, la véritable réussite de ce dossier.

Notre satisfaction est que notre alphabet soit enfin reconnu de manière internationale par des mesures pratiques capables de le sortir de son isolement.

Le codage de Tifinaghe occupe la rangée 2D dans le PMB. Il contient 54 caractères, un caractère diacritique et 25 emplacements vides pour d'éventuels réarrangements futurs. Le tableau suivant rassemble les caractères en question avec leurs points de codes en hexadécimal avec $x=0$:

Clé	
	Tifinaghe Ircam de base
	Tifinaghe Ircam étendu
	Autres lettres néotifinaghes
	Lettres touarègues modernes attestées
	Réservé pour un codage ultérieur

	2D3x	2D4x	2D5x	2D6x	2D7x
0	◦	⊙	≠	Δ	
1	⊖	∅	!	⊔	
2	⊕	⋮	∂	∫	
3	⊗	∠	∴	⌘	
4	⊗	∩	○	↑	
5	⊗	⊗	⊙	⌘	
6	⊔	∴	⊔		
7	∧	∩	∴		
8	∨	∴	∴		
9	⊕	⊗	⊙		
A	⊖	∩	⊙		
B	∴	⊗	⊙		
C	⊔	≠	+		
D	⊔	∥	⊗		
E	∴	⊔	⊙		
F	⊗	∩	⊕	u	

G = 00
P = 00

N° hexa	Glyphe	Nom
2D30	ⵏ	LETTRE TIFINAGHE YA
2D31	ⵍ	LETTRE TIFINAGHE YAB
2D32	ⵍⵎ	LETTRE TIFINAGHE YAB SPIRANT
2D33	ⵍⵏ	LETTRE TIFINAGHE YAG
2D34	ⵍⵏⵎ	LETTRE TIFINAGHE YAG SPIRANT
2D35	ⵍⵏⵔ	LETTRE TIFINAGHE YADJ KABYLE
2D36	ⵍⵏⵔⵎ	LETTRE TIFINAGHE YADJ
2D37	ⵍⵏⵕ	LETTRE TIFINAGHE YAD
2D38	ⵍⵏⵕⵎ	LETTRE TIFINAGHE YAD SPIRANT (yadh)
2D39	ⵍⵏⵕⵎⵔ	LETTRE TIFINAGHE YADD
2D3A	ⵍⵏⵕⵎⵔⵎ	LETTRE TIFINAGHE YADD SPIRANT
2D3B	ⵍⵏⵕⵎⵔⵎⵔ	LETTRE TIFINAGHE YEY
2D3C	ⵍⵏⵕⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAF
2D3D	ⵍⵏⵕⵎⵔⵎⵔⵎⵔ	LETTRE TIFINAGHE YAK
2D3E	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔ	LETTRE TIFINAGHE YAK TOUAREG
2D3F	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAK SPIRANT
2D40	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAH (yah touareg)
2D41	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔ	LETTRE TIFINAGHE YAH KABYLE
2D42	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔ	LETTRE TIFINAGHE YAH TOUAREG
2D43	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAHH
2D44	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YA'
2D45	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAKH
2D46	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAKH TOUAREG (quatre-points en carré touareg)
2D47	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAQ
2D48	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAQ TOUAREG
2D49	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YI
2D4A	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAJ
2D4B	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAJ DE L'AHAGGAR
2D4C	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAJ TOUAREG
2D4D	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAL
2D4E	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAM
2D4F	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAN
2D50	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YAGN TOUAREG
2D51	ⵍⵏⵕⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎⵔⵎ	LETTRE TIFINAGHE YANG TOUAREG

2D52	ⵝ	LETTRE TIFINAGHE YAP
2D53	ⵞ	LETTRE TIFINAGHE YOU (yaw touareg)
2D54	ⵟ	LETTRE TIFINAGHE YAR
2D55	ⵠ	LETTRE TIFINAGHE YARR
2D56	ⵡ	LETTRE TIFINAGHE YAGH
2D57	ⵢ	LETTRE TIFINAGHE YAGH TOUAREG
2D58	ⵣ	LETTRE TIFINAGHE YAGH DE L'AIR (yadj de l'Adrar, cinq-points en quinconce touareg)
2D59	ⵤ	LETTRE TIFINAGHE YAS
2D5A	ⵥ	LETTRE TIFINAGHE YASS
2D5B	ⵦ	LETTRE TIFINAGHE YACH
2D5C	ⵧ	LETTRE TIFINAGHE YAT
2D5D	⵨	LETTRE TIFINAGHE YAT SPIRANT (yath)
2D5E	⵩	LETTRE TIFINAGHE YATCH
2D5F	⵪	LETTRE TIFINAGHE YATT
2D60	⵫	LETTRE TIFINAGHE YAV
2D61	⵬	LETTRE TIFINAGHE YAW
2D62	⵭	LETTRE TIFINAGHE YAY
2D63	⵮	LETTRE TIFINAGHE YAZ
2D64	ⵯ	LETTRE TIFINAGHE YAZ TAWELLEMET (yaz harpon)
2D65	⵰	LETTRE TIFINAGHE YAZZ
2D66		(Cette position ne doit pas être utilisée)
2D67		(Cette position ne doit pas être utilisée)
2D68		(Cette position ne doit pas être utilisée)
2D69		(Cette position ne doit pas être utilisée)
2D6A		(Cette position ne doit pas être utilisée)
2D6B		(Cette position ne doit pas être utilisée)
2D6C		(Cette position ne doit pas être utilisée)
2D6D		(Cette position ne doit pas être utilisée)
2D6E		(Cette position ne doit pas être utilisée)
2D6F	⵱	LETTRE MODIFICATIVE TIFINAGHE DE LABIO-VÉLARISATION (amabat)
2D70		(Cette position ne doit pas être utilisée)
2D71		(Cette position ne doit pas être utilisée)
2D72		(Cette position ne doit pas être utilisée)
2D73		(Cette position ne doit pas être utilisée)
2D74		(Cette position ne doit pas être utilisée)

2D75		(Cette position ne doit pas être utilisée)
2D76		(Cette position ne doit pas être utilisée)
2D77		(Cette position ne doit pas être utilisée)
2D78		(Cette position ne doit pas être utilisée)
2D79		(Cette position ne doit pas être utilisée)
2D7A		(Cette position ne doit pas être utilisée)
2D7B		(Cette position ne doit pas être utilisée)
2D7C		(Cette position ne doit pas être utilisée)
2D7D		(Cette position ne doit pas être utilisée)
2D7E		(Cette position ne doit pas être utilisée)
2D7F		(Cette position ne doit pas être utilisée)

RÉFÉRENCES

- [1] H. MERTENS, *Éléments de la théorie de l'information et du codage*, Service de télécommunications, Faculté des sciences appliquées, université de liège, octobre 1987.
- [2] Jan Michel BERNARD et all., *De la logique câblée aux microprocesseurs*, édition Eyrolles, 1979, collection technique et scientifique des télécommunications.
- [3] Bernard DESGRAUPES, *Passeport pour Unicode*, Éditions Vuibert Informatique.
- [4] Yannis YARALAMBOUS, *Fontes et codage*, O'Reilly, 2004.
- [5] Site officiel de W3C, www.w3.org.
- [6] Site Web : UNICODE : <http://www.Unicode.org> ; ISO : <http://www.iso.org>
- [7] Patrick ANDRIES, *Unicode en français* : <http://cooptel.qc.ca> ; <http://pages.infinit.net/hapax> ;
- [8] *Correspondance à propos du codage avec Alain Labonté, expert et rédacteur des normes auprès de l'ISO*, 29 octobre 2003.
- [9] François YERGEAU : *Compte-rendu des réunions de Toronto, correspondance*, 21-25 juin à Toronto, Canada
- [10] *Compte rendu d'une réunion tenue à l'IRCAM (Rabat) le 19 février 2004, au sujet du codage des tfinagh dans l'ISO/CEI 10646 et Unicode*.
- [11] A. LA BONTÉ, P. ANDRIES, F. Yergeau, *Correspondance avec rapport sur les dispositions du clavier*, 6 avril 2004.
- [12] P. ANDRIES, *Rapport des discussions avec le ceisic sur le codage de la labialisation et impact sur le tri*, 2 mars 2004.
- [13] *Proposition de codage déposée auprès des instances internationales ISO et Unicode* ; 6 juin 2004, <http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2739.pdf>.
- [14] *Rapport d'activité du CEISIC, Inγmism n usinag, Bulletin d'information de l'Institut royal de la culture amazighe*, semestriel n° 3 et 4, Mars 2005, pp. 67-69.