Études et Documents Berbères, 18, 2000 : pp. 87-100

UNICODE ET ISO/IEC 10646

Introduction au codage informatique de l'écriture arabe et des alphabets berbères (latin et tifinagh)

par

Rachid Zghibi [1]

I. LE CODAGE INFORMATIQUE DES CARACTÈRES

Dans le domaine de la normalisation, les qualificatifs *« de jure »* et *« de facto »* sont utilisés comme premier critère de classification des normes [2]. Une norme de jure (ou formelle) désigne une norme qui a été développée par un organisme légalement constitué et mandaté pour élaborer des normes. Au niveau international, on trouve l'ISO (Organisation Mondiale de Normalisation), l'IFLA (International Federation of Library Associations and Institutions), et la CEI (Commission Électronique Internationale). Au niveau national, chaque pays a son organisme de normalisation. Par exemple en France on trouve l'AFNOR (Association Française de Normalisation), aux États-Unis on trouve l'ANSI (American National Standards Institute), au Royaume-Unie on trouve BSI (British Standards Institution), etc. Selon cette définition, toutes les normes élaborées par l'ISO ainsi que par tous les organismes nationaux et internationaux de normalisation sont des normes de jure.

Une norme de facto (standard) désigne une norme qui a été élaborée par

1. Doctorant au département Documentation à l'université Paris 8, chercheur à la Maison de Sciences de l'Homme (MSH) et membre au groupe de recherche SEM@TICE : un groupe d'enseignants et chercheurs francophones de l'AUF (Agence Universitaire de la Francophonie) engagés dans l'e-sémantique et l'e-learning.

2. CEVEIL (1995), *Normalisation et internationalisation : inventaire et prospectives des normes clefs pour le traitement informatique du français* (Rapport Bourbeau-Pinard) : http://www.ceveil.qc.ca/Normes.html

une organisation autre qu'un organisme formel de normalisation. Une norme de facto peut provenir de plusieurs sources d'activités : un consortium de fabricants d'ordinateurs, un consortium de vendeurs, un groupe d'utilisateurs, une association professionnelle, etc. *« Norme est définie par la publication (édition et prescription) d'un ensemble de règles, d'opérations, de protocoles, de caractéristiques généralement édifiées sous l'égide d'organismes institutionnels, comme l'ISO au niveau international et l'AFNOR en France. Standard renvoi à « standard de fait » et peut se définir comme l'ensemble de règles, d'opérations, de protocoles, de caractéristiques couramment employés et consacrés par l'usage »*[3].

Dans le domaine de codage informatique de caractères, il existe plusieurs normes de « jure » et de « facto ». Les jeux de caractères existants sont répertoriés en détail dans le standard Internet RFC 1345[4].

Un jeu de caractères respecte généralement certains critères :

Nom : un jeu de caractères est désigné sous divers appellations : codage de caractères, répertoire codé de caractères et page de code. Il est toujours nommé afin que le système de traitement ou de réception puisse utiliser la bonne table. Exemples de jeu de caractères : ISO/IEC 8859-1, ISO/IEC 8859-2, Page de code 1256, Unicode 4.0.0, etc. ;

Taille : la taille d'un jeu de caractères est exprimée par le nombre de bits[5] utilisé pour représenter un caractère. Le nombre de bits détermine également le nombre de caractères coder par le jeu de caractères. Actuellement, la majorité des jeux de caractères sont à 7 bits (ASCII, ISO/IEC 9036, etc.), à 8 bits (ISO/IEC 8859-n, pages de codes Dos et Windows, IBM EBCDIC, etc.), à 16 bits (UCS-2, Unicode, KSC 5601 (coréen), GB2312 (chinois simplifié), Big5 (chinois traditionnel)) et 32 bits (UCS-4, UTF 32).

Caractères : on trouve par exemple pour l'alphabet latin : les 10 chiffres, les 26 lettres (majuscules et minuscules), des signes de ponctuation ou des opérateurs, des caractères de contrôle. Tous les jeux de caractères distinguent le caractère de glyphe. Par définition, un caractère *« est une unité d'information utilisée pour coder du texte, alors qu'un glyphe est une forme géométrique (une collection homogène de telles formes constitue une police) utilisée pour présenter un texte. Le processus de présentation nécessite une application (pas nécessairement bi-univoque) des caractères vers des glyphes »*[6]. Les caractères forment les plus petites unités distinctives et significa-

---

3. Gabriel Gallezot et Franck Samson [...], « Normes et standards dans le processus de traitement du document numérique en biologie moléculaire », *Revue SOLARIS*, décembre 1999/janvier 2000. URL : http ://www.info.unicaen.fr/bnum/jelec/Solaris/d06/index.html

4. Simonsen, K., RFC 1345 : *Character mnemonics et character sets*, juin 1992, 103 p. URL : http ://www.ietf.org/rfc/rfc1345

5. Bit vient de la contraction de Binary Digit.

6. Jacques André, *op. cit.*, p. 6.

tives d'une langue écrite et les glyphes représentent les différentes formes qu'un caractère peut prendre.

Un caractère est représenté par un numéro qui ne réside qu'en mémoire ou sur disque. Par contre, le glyphe apparaît à l'écran ou sur papier et représente une forme particulière d'un ou plusieurs caractères. Différents scénarios existent : un seul glyphe peut représenter un seul ou plusieurs caractères ou de multiples glyphes peuvent représenter un seul caractère. Pour certaines écritures, comme l'arabe, le nombre de glyphes nécessaires à l'affichage peut être nettement plus important que le nombre des caractères qui codent les unités de base de cette écriture. Un répertoire de glyphes constitue une police. Une norme de codage porte uniquement sur les codes de caractères et non sur les glyphes. Le tableau suivante illustre les différences entre caractère et glyphe :

Différences entre caractère et glyphe

Glyphes	Caractères Unicode	Remarques
A **A** A A A A A A **A**	U+0041 LETTRE MAJUSCULE LATINE A	Plusieurs glyphes pour représenter un seul caractère.
? ? ? ?	U+062E LETTRE ARABE KHA	
fi	U+0066 LETTRE MINUSCULE LATINE F + U+0069 LETTRE MINUSCULE LATINE I	Un glyphe pour représenter deux caractères

Traitement : le codage des chiffres et des lettres doit être conçu pour faciliter le traitement. On doit par exemple pouvoir facilement faire le tri, le changement majuscule/minuscule sur des caractères. Donc, puisque au sens lexicographique « AB », le code du caractère « A » doit être plus petit (au sens arithmétique, puisque c'est un nombre) que le code du caractère « B ». Le principe est le même pour les chiffres. Les relations d'ordre entre les chiffres doivent être respectées pour les codes (5 étant plus grand que 2 ; le code du caractère 5 doit être plus grand que le code du caractère 2).

AU-DELÀ DU CODAGE DES CARACTÈRES : LES PROBLÈMES DE TRANSMISSION DES DONNÉES

Fondamentalement, les ordinateurs ne comprennent que des instructions fondées sur des codes numériques. Par conséquent, les caractères (les lettres, les chiffres et les autres symboles) sont enregistrés sur un disque dur ou sur n'importe quels supports de stockage sous formes des nombres qui sont utilisés par l'ordinateur pour afficher ou échanger les caractères. Par exemple, dans la

norme ISO/IEC 8859-1, le caractère A (LETTRE LATINE A MAJUSCULE) a le numéro 65 et le caractère a (LETTRE LATINE a MINUSCULE) a le numéro 97.

À l'ère actuelle, des centaines de systèmes de codage de caractères ont été créés. Ces systèmes de codage sont souvent incompatibles entre eux. Ainsi, deux systèmes peuvent utiliser le même code numérique (numéro) pour deux caractères différents ou utiliser différents codes pour le même caractère. Par exemple, la lettre ? (p) de l'alphabet persan a la valeur numérique 129 dans le tableau des codes du jeu des caractères CP-1256 et elle a la valeur numérique 243 dans le tableau des codes du jeu de caractères CP-Arabic Mac. Le problème de transfert des données entre systèmes ou des migrations d'un système à l'autre se pose alors. Ce problème n'est donc pas de carence, mais bien de surabondance. Pour qu'un groupe d'utilisateurs qui partagent la même écriture puisse échanger, il faut :

– qu'ils s'entendent sur un petit nombre (un seul, si possible) de codages ;

– que leurs logiciels reconnaissent et puissent traiter tous ces codages.

Par conséquent, la multitude des normes et des standards de codage de caractères complique énormément la tâche des concepteurs de logiciels internationaux et des utilisateurs de plusieurs langues surtout dans le cas où ces langues ne pourraient être représentées par un seul codage. *« Les traitements de texte multilingues doivent jongler à la fois avec les différentes normes de codage et les polices associées »*[7]. À titre d'exemple, pour l'écriture arabe, on dispose de plusieurs jeux de caractères : ISO 8859-6, ISO 9036, ASMO 449, MS Arabic Dos Code Page 708 (ASMO 708), MS Windows Arabic Code Page 1256, Arabic Mac Code Page, Arabic Windows 3X Code Page, Code Page 864 Dos Arabic, etc.

II. LA NORME ISO/IEC 10646 ET LE STANDARD UNICODE : VERS UNE SOLUTION UNIVERSELLE

Aperçu historique

En 1984, le JTC1/SC2 de l'ISO/IEC a constitué le Groupe de Travail n° 2 (WG2) qui avait comme objectif d'élaborer une norme universelle de codage de caractères graphiques des langues écrites du monde. *« En fait, ce nouveau mode de codage permet de supporter simultanément toutes les écritures du*

7. Claude De Loupy, « Multilinguisme et document numérique : la dimension technique à l'épreuve du codage des caractères », *Revue SOLARIS*, décembre 1999/janvier 2000. URL : http ://www.info.unicaen.fr/bnum/jelec/Solaris/d06/index.html

monde, mais aussi les diverses notations musicales, des écritures ésotériques, les notations chorégraphiques présentes ou à venir, les langues de signes, l'étude des langues anciennes »[8]. En 1990, ce groupe de travail a développé une première version de la norme ISO/IEC 10646. Elle définit un ensemble universel de codage des caractères sur plusieurs octets (Universal Multiple Octet Coded Character Set) plus précisément sur quatre octets. Ce jeu de caractères universel est désigné par le sigle UCS, pour *Universal Character Set*. Le codage ISO/IEC 10646 s'effectue sur quatre octets, qui sont appelés en commençant par le moins significatif, la cellule (octet C), la Ligne ou la Rangée (octet R), le Plan (octet P) et le Groupe (octet G). *« Cette forme canonique utilise donc quatre dimensions qui consiste en 128 groupes tridimensionnels ; chaque groupe consiste en 256 plan bidimensionnels et chaque plan à 256 lignes unidimensionnelles avec chacune 256 cellules »*[9].

La norme ISO/IEC 10646 définit deux formes de codage :

1. un codage sur quatre octets (32 bits) qui comprend 4 294 967 296 positions de code.

2. un codage sur deux octets, ou un seizet (16 bits), qui correspond au plan 0 du groupe 0 dit le Plan Multilingue de Base (BMP). Il comprend 65 535 positions de code.

La forme de codage sur 32 bits est connue sous le nom d'UCS-4 (jeu universel de caractères sur quatre octets). Dans l'UCS-4, le premier octet indique le numéro du groupe (G), le deuxième octet indique le numéro du plan (P), le troisième octet indique le numéro de la ligne ou la rangée (R) et enfin le quatrième octet indique le numéro de la cellule (C).

La forme de codage sur 16 bits porte le nom d'UCS-2 (jeu universel de caractères sur deux octets). Les numéros des codes de 0 à 65 535 décimal (0000 à FFFF hexadécimal) peuvent être représentés à l'aide d'une valeur de code de caractère à 16 bits. Ces codes sont affectés au BMP (Basic Multilingual Plane, Plan de Base Multilingue). Les 65 536 positions des codes du BMP sont représentées dans 256 lignes dont chacune est composée de 256 cellules. Dans l'espace de codage de l'UCS-2 le premier octet indique le numéro de la ligne (G)et le deuxième octet indique le numéro de la cellule (C).

Parallèlement aux travaux de l'ISO, le consortium Unicode s'est constitué

---

8. Henri Hudrisier, *Le document numérique normalisé XML, TEI, Unicode : nouvelles automatisations de l'intelligence et opportunités de prospérité : enjeux pédagogiques, nouvelles organisations de la recherche en ingénierie du document*, Colloque AUPELF : Universités virtuelles, vers un enseignement égalitaire, Edmundston, 26, 30 août 1999.

9. ISO, *International Standard ISO/IEC 10646-1 : information technology-universal multiple-octet coded character set (UCS) : part1 architecture and basic multilingual plane*, Genève : ISO, 1993, p. 3.

en 1991 sous le nom de Unicode inc. Le consortium Unicode est composé de la plupart des sociétés informatiques les plus importantes dans le monde de l'informatique à savoir : IBM, MICROSOFT, LOTUS, SUN, XEROX, ADOBE, ORACLE, NOVELL, APPLE, COMPAQ, etc. [10]

Le consortium Unicode propose une méthode uniforme pour le codage et l'identification des caractères basée sur un code fixe à 16 bits. Le nouveau système a pour but de permettre l'échange, le traitement et la visualisation des caractères de la majorité des systèmes d'écritures alphabétiques, syllabiques et idéographiques en usage dans le monde. Le consortium Unicode a nommé son standard ***Unicode*** [11]. La première version du standard a été publiée en 1991.

En 1992, à la suite de négociations officielles sur une convergence des deux codes, les deux répertoires ont été fusionnés. La fusion des deux codes consistait à faire correspondre les valeurs numériques des caractères identiques et d'augmenter le répertoire de caractères présents dans la norme ISO/IEC 10646 mais qui sont absents dans le standard Unicode. Depuis lors, le WG2 de l'ISO/IEC et le comité technique d'Unicode (UTC : le groupe de travail au sein du consortium qui est chargé de la création et de la mise a jour du standard) travaillent conjointement afin d'assurer la compatibilité entre les deux jeux de caractères. En mai 1993, la norme internationale ISO/IEC 10646-1 : 1993, Technologies de l'information Jeu universel de caractères codés sur plusieurs octets (JUC) – Partie 1 : architecture et plan multilingue de base a été publiée. La version 1.1 du standard Unicode est parfaitement identique à la norme ISO/IEC 10646 et tient compte des caractères supplémentaires introduits par la norme ISO.

La version 4.0.0 (version majeure [12]) est la dernière version du standard Unicode. Elle est publiée en 2003 et elle est rigoureusement identique à la norme ISO/IEC 10646 : 2003. Par rapport à la version 3.2 du standard Unicode, l'actuelle version ajoute 1 226 nouveaux caractères pour atteindre un total de 96 248 caractères graphiques : 50 635 caractères graphiques sont codés au niveau du BMP et 45 613 caractères graphiques sont codés dans les plans supplémentaires. Concernant les nouveaux caractères : 452 caractères

10. Pour la liste complète des membres du Consortium Unicode voir : http://www.unicode.org/unicode/consortium/memblogo.html

11. http://www.unicode.org

12. On distingue trois types de versions du standard Unicode : la version majeure, la version mineure et la mise à jour. La version majeure consiste aux additions importantes au standard. Elle est publiée sous la forme d'un livre. La version mineure correspond aux ajouts des caractères ou aux certains changements normatifs plus conséquents. Elle est publiée sous la forme d'un rapport technique sur le site Internet d'Unicode. La mise à jour correspond aux autres changements aux sections informatives ou aux sections normatives du standard. Ces changements entraînent la production d'un nouveau fichier UnicodeData.txt et d'autres fichiers connexes de la base de données de caractères Unicode. Pour plus d'informations, voir http://www.unicode.org/Unicode/standard/versions/

sont codés dans le BMP et 774 dans les plans supplémentaires. Ils concernent les blocs de caractères latins, cyrilliques, l'écriture thaïe ainsi que plusieurs écritures minoritaires et historiques. Le tableau suivant compare le nombre des caractères codés au niveau du BMP dans les différentes versions du standard Unicode et la norme ISO/IEC 10646[13] :

III. ÉCRITURE ARABE

« L'arabe fait partie des langues chamito-sémitique et plus précisément, à l'intérieur de cet ensemble, du groupe des langues sémitiques, attesté depuis la plus haute Antiquité et constituant, jusqu'à aujourd'hui, l'une des grandes familles de l'expression humaine »[14]. Le groupe des langues sémitiques se divise en sémitique oriental (avec l'akkadien), sémitique occidental ou du nord (cananéen, phénicien, hébreu, araméen, ce dernier est connu notamment sous sa forme syriaque), sémitique méridional. L'arabe relève de ce dernier groupe, avec l'éthiopien et le sud-arabique.

Caractéristiques

L'écriture arabe est une écriture fondamentalement consonantique. Pour préciser la prononciation, 3 voyelles brèves et 7 signes orthographiques s'ajoutent aux consonnes et se positionnent soit au-dessous soit au-dessus des consonnes. Comme les autres écritures sémitiques (hébreu, syriaque), l'arabe s'écrit toujours de droite à gauche.

1 - *Consonnes*

l'alphabet arabe est composé de 29 consonnes (voir tableau n° 1).

2 - *Les voyelles brèves*

Les trois voyelles brèves sont les suivantes :

– **fatha** « ? U + 064E[15] » (elle surmonte la consonne et se prononce comme un a en français),

13. http://www.unicode.org/versions/Unicode4.0.0/

14. Andre Miquel, « Les gens du Dad », *Les Arabes, du message à l'histoire*, Paris, Librairie Arthème Foyard, 1995, p. 54.

15. Dans le standard Unicode, chaque caractère est identifié soit par sa valeur numérique (son numéro de 0 à 65 535 dans le BMP) soit par sa valeur hexadécimale (le numéro de la rangée et le numéro de la cellule). Le numéro de la cellule comprend le numéro du colonne et le numéro de la ligne. Les deux types de valeurs (numérique et hexadécimale) sont toujours précédées par U + . Par exemple la voyelle fatha a la valeur numérique U + 1614 et a la valeur hexadécimale U + 064E

– **damma** « ? U + 064F » (elle surmonte la consonne et se prononce comme un u en français),

– **kasra** « ? U + 0650 » (elle se note au-dessous de la consonne et se prononce comme un i en français).

3 - *Les signes orthographiques*

Les sept signes orthographiques sont les suivants :

– **sukun** « ? U + 0652 » : ce signe indique qu'une consonne n'est pas suivie (ou mue) par une voyelle. Il est noté toujours au-dessus de la consonne ;

– Les trois signes de tanwin : lorsque les trois voyelles brèves (fatha, kasra et damma) sont doublées, elles prennent un son nasal, comme si elles étaient suivies de "n" et on les prononce respectivement :

✓ an « ? U + 064B » pour la **fathatan**

✓ in « ? U + 064D » pour la **kasratan**

✓ un « ? U + 064C » pour la **dammatan**

À l'instar des voyelles brèves, la fathatan et la dammatan se positionnent toujours au-dessus de la consonne contrairement à la kasratan qui ne se place qu'au-dessous de la consonne ;

– **chadda** « ? U + 0651 » comme dans le français "immédiatement", l'arabe peut renforcer une consonne quelconque. Ce renforcement est indiqué à l'aide d'un signe nommé *chadda* ou *tachdid* (intensité) ;

– **wasla** « ? U + 0671 » : quand la voyelle d'un *alif* au commencement d'un mot doit être absorbée par la dernière voyelle du mot qui précède, on en indique l'élision par le signe wasla placé au-dessus de l'*alif ;*

– **madda** « ? U + 0653 » : le *madda* (prolongation) se place sur l'*alif* pour indiquer que cette lettre tient lieu de deux *alifs* consécutifs ou qu'elle ne doit pas porter le *hamza*. Ce signe de contraction a la forme d'un *alif* horizontal. Le *madda* surmonte aussi les groupes de lettres exprimant une abréviation. C'est ainsi qu'on le trouve sur les lettres énigmatiques qui commencent certains chapitres du Coran, et sur celles qui représentent en abrégé quelques formules.

(06 est le numéro de la rangée, 4 est le numéro du colonne et E est le numéro de la ligne). Les caractères Unicode utilisés dans l'article sont tous identifiés par leurs valeurs hexadécimales.

16. Pour chaque lettre arabe, nous avons donné sa valeur hexadécimale selon le standard Unicode et la norme ISO/IEC 10 646 et son équivalent en caractère latin conformément à la norme ISO de translittération des caractères arabes en caractères latins : ISO 233, décembre 1966.

N°	Caractères arabes[16]	Valeurs Unicode	Translittération en caractères latins	N°	Caractères arabes	Valeurs Unicode	Translittération en caractères latins
1	ء	U+0621	ʾ	16	ض	U+0636	ḍ
2	ا	U+0627	a	17	ط	U+0637	ṭ
3	ب	U+0628	b	18	ظ	U+0638	ẓ
4	ت	U+062A	t	19	ع	U+0639	ʿ
5	ث	U+062B	t̲	20	غ	U+063A	ḡ
6	ج	U+062C	ǧ	21	ف	U+0641	f
7	ح	U+062D	ḥ	22	ق	U+0642	q
8	خ	U+062E	h̲	23	ك	U+0643	k
9	د	U+062F	d	24	ل	U+0644	l
10	ذ	U+0630	d̲	25	م	U+0645	m
11	ر	U+0631	r	26	ن	U+0646	n
12	ز	U+0632	z	27	ه	U+0647	h
13	س	U+0633	s	28	و	U+0648	w
14	ش	U+0634	š	29	ي	U+064A	y
15	ص	U+0635	ṣ				

Tableau 1 : Les consonnes arabes

Unicode ET ISO/IEC 10646 : le codage des caractères arabes dans le BMP

Dans la version 4.0.0, le standard Unicode et la norme ISO/CEI 10646 réserve 1 026 codes pour coder l'écriture arabe de base ainsi que les différentes écritures dérivées de l'arabe de base. Les caractères sont regroupés en blocs de caractères au niveau de la rangée 6, FB, FC, FD et une partie de la rangée FE du Plan Multilingue de Base (BMP). [17]

Rangée 6

La rangée 6 est destinée à coder les caractères arabes de base et étendu. Elle code 227 caractères qui se présentent comme suit :

– les lettres arabes de base ;

– les lettres arabes modifiées qui sont utilisées par les alphabets qui sont

17. Pour plus d'informations sur le traitement de l'écriture arabe par le standard Unicode et la norme ISO/IEC 10646, voir : Zghibi Rachid, « Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/IEC 10646 », *Document Numérique*, Volume 6, n° 3, 2002, pp. 155-182

dérivés de l'alphabet arabe de base comme l'alphabet persan, urdu, pashto, etc. ;

– les signes de vocalisation (voyelles brèves et signes orthographiques) ;

– les signes arabes de ponctuation tels que la virgule arabe (؟ : U+060C), le point-virgule arabe (؟ : U+061B), le point d'interrogation arabe (؟ : U + 061F) ;

– les chiffres (arabes, hindous et les chiffres utilisées dans l'écriture Urdu) ;

– les signes utilisés dans le texte coranique tels que le symbole qui indique la fin d'un verset dans le Coran (06DD), etc.

– les signes de vocalisation additionnels qui sont utilisés dans les écritures dérivées de l'arabe de base, etc.

Rangée FB

La rangée FB (positions FB50 à FBFF) est destinée à coder les différentes formes contextuelles des caractères des alphabets dérivés de l'arabe de base (isolées, initiales, médianes et finales). La rangée FB code 143 formes.

Rangées FC et FD

Les rangées FC (positions FC00 à FCFF) et FD (positions FD00 à FD7F) sont réservées exclusivement à coder un grand nombre de ligatures esthétiques à deux ou à trois lettres initiales et finales, ainsi que les ligatures qui forment des mots en entier ou un groupement de mots. Sont codés également, les différentes formes de combinaison de voyelles avec les signes orthographiques ainsi que d'autres signes. Ces deux rangées codent ensemble 516 caractères.

Rangée FE

Une partie de la rangée FE (positions FE70 à FEFF) est destinée à coder les variantes contextuelles de lettres arabes de base et de la ligature linguistique LAM-ALIF (؟). On trouve également les différentes positions qui peuvent avoir les voyelles brèves (Fatha, Kasra, Damma) et les signes orthographiques au-dessous et au-dessus des consonnes. La partie arabe de la rangée FE comprend 140 codes.

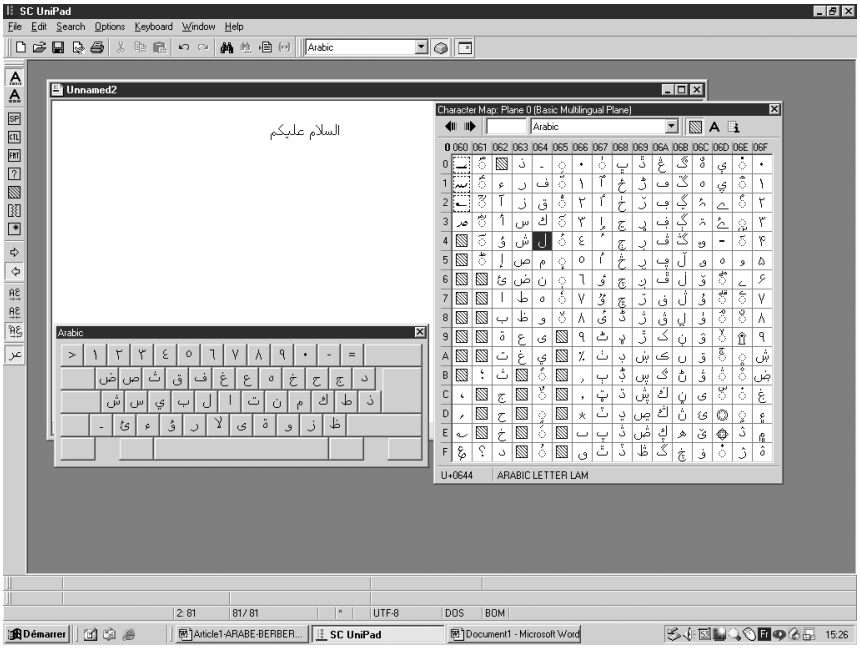


Figure 1 : Le bloc de caractères arabes de base et étendu de la rangée 6 sous l'éditeur de texte UniPad.

IV. LE BERBÈRE

Le berbère appartient au groupe des langues chamito-sémitiques. Cependant, elle n'est pas une langue unique mais plusieurs dialectes qui constituent en fait de véritables langues. Sur le plan graphique, trois graphies sont en usage actuellement pour noter les langues berbères : les caractères arabes, les caractères latins et les tifinagh. *« Les tifinagh sont l'écriture du sentiment et de l'identité, les caractères latins celle de la rigueur scientifique et du modernisme, les caractères arabes celle de la tradition et, tout bien considéré, du bon sens des langues plus ou moins autonomes »*[18]. Dans ce chapitre on va étudier uniquement les systèmes d'écritures tifinagh et les systèmes d'écritures latins.

Les tifinagh :

« Tifinagh (féminin pluriel) a pour singulier Tafineq « caractère, signe, lettre

18. Olivier Durand, « Promotion du berbère : problèmes de standardisation et d'orthographe : expériences européennes », *Études et documents berbères*, n° 11, 1994, p. 8.

d'écriture », mot peu usité. Par tifinagh, les Touaregs désignent l'ensemble des caractères de leur écriture » [19].

Cependant, il convient de distinguer les *tifinagh authentiques*, originelles (en usage chez les Touaregs sahariens d'Algérie, de le Libye, du Mali, du Niger, au Burkina-Faso et dans la diaspora touarègue vivant au Nigeria, au Ghana, au Côte d'Ivoire, au Soudan, au Tchad) des *néo-tifinagh* (en usage en Occident et en régions berbérophones).

Les tifinagh originelles sont héritées des Anciens Touaregs. D'après Mohamed Aghali-Zakara [20], il existe une douzaine voire plus d'alphabets tifinagh selon les parlers. Les signes constituant ces alphabets comportent des variants et des invariants. Cependant, quelque soit le parler, seules les consonnes sont écrites sans marque de gémination ou de tension. La voyelle « a » représentée par le signe « . » n'est notée qu'en fin de mot ou de texte. Le même signe peut également avoir la valeur de la voyelle « i » et « u ». Depuis les années 1960, les tifinagh d'origine sont employées dans les journaux locaux et dans les centres d'alphabétisation des adultes au Mali et au Niger.

Les néo-tifinaghs désignent les systèmes d'écriture développés pour représenter les parlers berbères du Maghreb. La première variante fut celle proposée à la fin des années 1960 par l'Académie berbère (AB) [21]. L'objectif de l'Académie était d'adapter les tifinagh traditionnelles à la notation du kabyle, dialecte qui représente de nombreuses particularités phonétiques. Pour se faire, l'alphabet Ahaggar a été retenu comme base pour le développement de son nouveau alphabet. Il comporte 37 signes dont 21 signes sont nouveaux par rapport à l'alphabet de l'Ahaggar. « *Il s'agit d'un alphabet qui veut à la fois s'adapter à la* phonologie kabyle, remplacer les caractères jugés peu pratiques, particulièrement ceux comportant des points, et d'adjoindre des caractères vocaliques » [22]. Le nouveau alphabet est largement diffusé au Maroc et en Algérie surtout en Kabylie.

Plusieurs systèmes ont été développés par la suite avec le but d'améliorer les quelques imperfections du système de l'Académie berbère. On cite notamment le système proposé par Salem Chaker [23] qui représente une variante néo-tifinagh développée par l'Institut royal de la culture amazighe du Maroc (Ircam).

Les caractères latins :

Les premières tentatives de la notation du berbère avec l'alphabet latin

19. Mohamed Aghali-Zakara, « Graphies berbères et dilemme de diffusion : interaction des alphabets latin, ajami et tifinagh », *Études et documents berbères*, n° 11, 1994 p. 110.

20. Mohamed Aghali-Zakara, *op. cit.* pp. 107-121.

21. C'est une association qui regroupe des intellectuels berbérophones. Elle a été créée à Paris en 1967.

22. Mohamed Aghali-Zakara, *ibid.* p. 117.

23. Voir l'article de Salem Chaker « Pour une notation usuelle à base tifinagh », *Études et documents berbères*, n° 11, 1994, pp. 31-42.

ont commencé depuis le XIX^e siècle notamment par les missionnaires, les militaires et divers chercheurs occidentaux. En dehors de toute harmonisation, plusieurs alphabets ont été définis avec beaucoup de divergence. Un travail d'harmonisation a commencé avec la conférence internationale de Bamako en 1966 et s'est poursuivi par celle de 1984. Les alphabets élaborés, puis améliorés pendant ces deux conférences, ont été officiellement adoptés au Mali et au Niger dans les services d'alphabétisation des adultes et dans les écoles primaires expérimentales en Touareg. L'alphabet latin adopté lors de la conférence de Bamako en juin 1984 est le suivant[24] :

a, ă, b, d, ?, e, ?, f, g, ?, h, ḥ, ç, i, j, k, l, ḷ, m, ?, ?, n, o, q, r, °, s, ṣ, š, t, ṭ, u, w, x, y, z, ẓ, ž

Cette liste reste ouverte pour noter tout autre phonème découvert ultérieurement. *« D'un point de vue pratique, le système des caractères latins est le plus attrayant. D'abord il est économique. Il ne compte qu'un nombre limité de caractères [...]. Ensuite, et si on vise le long terme, en fonction des rapports de force actuels et futurs, c'est le système qui est le plus utilisé comme support de vulgarisation du savoir, qu'il soit technologique ou autre. »*[25].

UNICODE ET ISO/IEC 10646 : le codage des écritures berbères

Dans cette section nous identifierons les codes réservés par le standard Unicode et la norme ISO/IEC 10646 pour coder les caractères latins et les caractères tifinagh. En tant qu'une proposition, les caractères tifinagh ne sont pas encore codés.

caractères latins :

Les caractères latins définis pendant les deux conférences internationales de Bamako en 1966 et en 1984 pour transcrire le berbère sont codés :

– au niveau des blocs de caractères *Latins de base* (Rangée 0, positions : U + 0020-U + 007E) et *Latin-1 Supplément* (rangée 0, positions : U + 00A0-U + 00FF) ;

– au niveau des blocs de caractères *Latins étendus A* (Rangée 1, positions 0100-017F) et *Latins étendus B* (rangée 1, positions 0100-017F) ;

– au niveau du bloc de caractères *IPA Extensions* (Rangée 2, positions 0250-02AF) ;

– au niveau du bloc de caractères *Latins étendus additionnels* (Rangée 1E, positions 1E00-1EFF) ;

Le tableau suivant identifie l'ensemble de caractères latins utilisés pour transcrire le berbère avec leurs positions exactes dans les différents blocs de caractères Unicode et ISO/IEC 10646.

---

24. Voir l'article de Mohamed Aghali-Zakara, *op.cit*, p. 109.

25. Mohyedine Benlakhdar, « Écrire le berbère : une nécessité scientifique ou pratique », *Études et documents berbères*, n° 11, 1994, p. 21.

Caractère berbère	Valeur Unicode (hexadécimale)	Caractère berbère	Valeur Unicode (hexadécimale)
a	U+0061	ɲ	U+0272 (IPA Extensions)
ă	U+0103 (Latin étendu-A)	ɳ	U+016E (Latin étendu-B)
b	U+0062	n	U+006E
d	U+0064	o	U+006F
ð	U+00F0 (Latin-1 Supplément)	q	U+0071
e	U+0065	r	U+0072
ɖ	U+018F (Latin étendu-B)	°	00B0 (Latin-1 Supplément)
f	U+0066	s	U+0073
g	U+0067	ṣ	U+1E63 (Latin étendu additionnel)
ɣ	U+0194 (Latin étendu-B), 0263 (IPA)	š	U+0161 (Latin étendu-A)
h	U+0068	t	U+0074
ḥ	U+1E25 (Latin étendu additionnel)	ṭ	U+1E6D (Latin étendu additionnel)
ç	U+00E7 (Latin-1 Supplément)	u	U+0075
i	U+0069	w	U+0077
j	U+006A	x	U+0078
k	U+006B	y	U+0079
l	U+006C	z	U+007A
l (avec un point au-dessous)	U+1E37 (Latin étendu additionnel)	ẓ	U+1E93 (Latin étendu additionnel)
m	U+006D	ž	U+017E (Latin étendu-A)

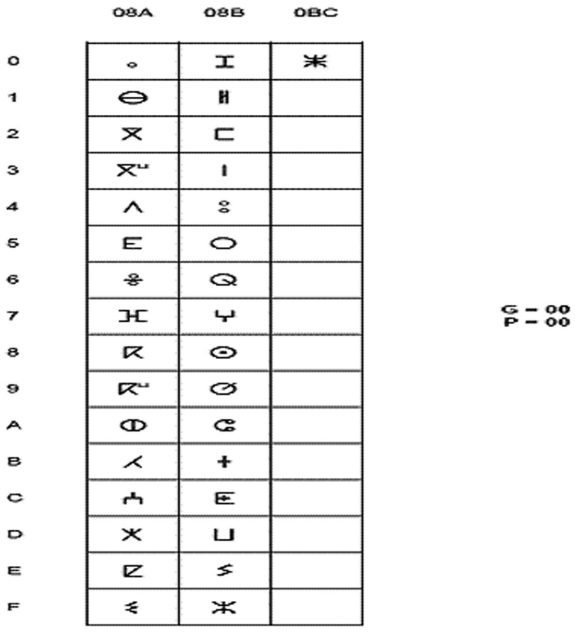
Les tifinagh : proposition d'addition de l'écriture tifinagh au répertoire de l'ISO/IEC 10646 et du standard Unicode

Le 26 janvier 2004, une proposition d'ajout de l'écriture tifinagh au répertoire de la norme ISO/IEC 10646 et du standard Unicode a été formulée par Patrick Andries, Français Yergeau et Alain LaBonté auprès de l'ISO/IEC/JTC1/SC2/GT2. La proposition reprend les symboles tifinagh retenus par l'Institut royal de la culture amazighe du Maroc (Ircam) qui sont au nombre de 33. Cette liste n'est pas définitive et elle pourrait être complétée par des nouveaux caractères.

C'est une écriture non cursive et non contextuelle. Elle s'écrit de gauche à droite et elle emploie les signes de ponctuation conventionnels qu'on retrouve

dans les écritures latines. Concernant les chiffres, il est préconisé d'utiliser les chiffres arabes occidentaux.

Ces 33 caractères seront codés dans le BMP au niveau de la rangée 8 (de la position 08A0 à la position 08C0). La figure suivante identifie l'ensemble de ces 33 caractères tifinagh en question et indique leurs emplacements respectifs dans la rangée 8 du BMP.



V. CONCLUSION

Sur la question du codage informatique du berbère, nous nous sommes limités aux deux systèmes graphiques utilisés : les caractères latins et les caractères tifinagh et nous avons montré sommairement comment ces caractères sont codés dans Unicode et dans la norme ISO/IEC 10646. Cependant, le berbère se note également en caractères arabes. À l'instar des alphabets dérivés de l'arabe de base tels que le persan, l'ourdou, le kurde sorani, le pashto, etc. il a fallu inventer de nouvelles lettres pour noter les phonèmes berbères qui n'ont pas d'équivalent dans la langue arabe. Dans le standard Unicode ces nouvelles lettres berbères sont codées dans la rangée 6 (caractères arabes de base et étendu). L'étude du codage informatique de ce troisième système graphique fera l'objet d'un prochain article.

Rachid Zghibi

RÉFÉRENCES

AGHALI-ZAKARA M., « Graphies berbères et dilemme de diffusion : interaction des alphabets latin, ajami et tifinagh », *Études et documents berbères*, n° 11, 1994 pp. 107-121.

ANDRÉ J. « Codage des caractères et multilinguisme : de l'ASCII à UNICODE et ISO/IEC 10646 », *Cahiers Gutemberg*, n° 20, juin 1995, pp. 1-50.

BENLAKHDAR M., « Écrire le berbère : une nécessité scientifique ou pratique », *Études et documents berbères*, n° 11, 1994, pp. 19-23.

CHAKER S., « Pour une notation usuelle à base tifinagh », *Études et documents berbères*, n° 11, 1994, pp. 31-42.

DE LOUPY C., « Multilinguisme et document numérique : la dimension technique à l'épreuve du codage des caractères », *Revue SOLARIS*, décembre 1999/janvier 2000. URL : http ://www.info.unicaen.fr/bnum/jelec/Solaris/d06/index.html

DURAND O., « Promotion du berbère : problèmes de standardisation et d'orthographe : expériences européennes », *Études et documents berbères*, n° 11, 1994, pp. 7-11.

FABRE, C., *De l'écrit au numérique : constituer, normaliser et exploiter les corpus électroniques*, Paris : Inter-Éditions, 1998, 320 p.

GALLEZOT G. et SAMSON F. « Normes et standards dans le processus de traitement du document numérique en biologie moléculaire », *Revue SOLARIS*, décembre 1999/janvier 2000. URL : http ://www.info.unicaen.fr/bnum/jelec/Solaris/d06/index.html

HUDRISIER H., *Le document numérique normalisé XML, TEI, Unicode : nouvelles automatisations de l'intelligence et opportunités de prospérité : enjeux pédagogiques, nouvelles organisations de la recherche en ingénierie du document*, Colloque AUPELF : Universités virtuelles, vers un enseignement égalitaire, Edmundston, 26, 30 août 1999.

ISO, *International Standard ISO/IEC 10646-1 : information technology-universal multiple-octet coded character set (UCS) : part1 architecture and basic multilingual plane*, Genève : ISO, 1993, p. 3.

MIQUEL A., « Les gens du Dad », *Les Arabes, du message à l'histoire*, Paris, Librairie Arthème Foyard, 1995, pp. 50-65.

SIMONSEN, K., RFC 1345 : *Character mnemonics et character sets*, juin 1992, 103 p. URL : http ://www.ietf.org/rfc/rfc1345 UNICODE. URL : http ://www.unicode.org

ZGHIBI, R., « Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/CEI 10646 », *Document Numérique*, Volume 6, n° 3, 2002, pp. 155-182.