Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/CEI 10646

Rachid Zghibi

Département documentation
Université Paris 8

Maison des Sciences de l'Homme Paris Nord
4, rue de la Croix Faron, Plaine Saint-Denis
F-93210 Saint-Denis

rachid_zghibi@yahoo.fr

RÉSUMÉ. *Le traitement de l'écriture arabe par le standard Unicode et la norme ISO/CEI 10646 constitue l'objet de cet article. Nous présentons d'abord les principaux problèmes liés au codage informatique de l'écriture arabe, puis nous montrons comment Unicode résout ces problèmes.*

ABSTRACT. *The present paper deals with the treatment of arabic script by the Unicode and ISO/CEI 10646. First, we shall outline the principal problems related to the encoding of arabic script. Then we shall show how Unicode resolves this problems.*

MOTS-CLÉS : *écriture arabe, Unicode, ISO/CEI 10646, codage.*

KEYWORDS: *arabic script, Unicode, ISO/CEI 10646, encoding.*

Document numérique. Volume 6 – n° 3-4/2002, pages 155 à 182

1. Écriture arabe

1.1.*Origine*

L'arabe fait partie des langues chamito-sémitique et plus précisément, à l'intérieur de cet ensemble, du groupe des langues sémitiques. Ce groupe se divise en sémitique oriental (avec l'akkadien), sémitique occidental ou du nord (cananéen, phénicien, hébreu, araméen) et sémitique méridional. L'arabe relève de ce dernier groupe avec l'éthiopien et le sud-arabique. À la différence d'autres nations, comme les anciens égyptiens, les babyloniens et les chinois dont les systèmes d'écriture remontent à des milliers d'années, l'écriture arabe n'est apparue qu'au VI^{e} siècle. L'origine de l'écriture arabe est assez obscure. En procédant à des études comparatives, certains historiens et chercheurs pensent que l'origine de l'écriture arabe est le syriaque en se basant sur plusieurs indices, à savoir :

- l'ordre primitif des lettres arabes est identique à celui des alphabets hébraïque et syriaque ;
- les formes de l'ancien alphabet arabe dit « coufique » est comparable à celle de « l'estranghelo », une forme de l'écriture syriaque, etc.

Cependant, la théorie la plus communément admise fait descendre celle-ci de l'écriture nabatéenne, l'ascendante directe de l'écriture araméenne ancienne qui est un rejeton de l'alphabet phénicien. De l'écriture nabatéenne, l'arabe a hérité ses vingt-deux lettres et l'emploi des ligatures pour réunir une lettre à la suivante. Par conséquent la plupart des lettres arabes peuvent revêtir quatre formes différentes : isolée, initiale, médiane et finale [4].

1.2. *Réformes*

L'arrivée de l'Islam a profondément marqué l'histoire de la langue et de l'écriture arabe. Le Coran, livre sacré recueillant la parole de Dieu mais aussi code juridique et moral, occupe d'emblée une place centrale dans la vie du croyant et de la communauté musulmane. Dans l'Islam des origines, le Coran était conservé dans la mémoire des compagnons du prophète Mahomet. Cependant, la mort de nombreux compagnons connaissant le Coran par cœur durant les batailles a poussé les premiers califes de l'Islam à collationner et à fixer le Coran sur un support scripturaire afin d'éviter sa disparition et sa manipulation.

À l'époque où le Coran a été fixé, l'écriture arabe posait beaucoup de problèmes. Les 22 consonnes nabatéennes sont insuffisantes pour écrire les 28 phonèmes arabes. À ce problème s'ajoute l'absence de points diacritiques et des signes vocaliques qui réduisait la graphie à un simple schéma consonantique. La réforme de l'écriture arabe devenait alors une affaire impérative et très urgente. C'est à l'époque du califat

Abbasside[1] que plusieurs réformes ont été accomplies par des pionniers de la langue et de la grammaire arabe.

Ajout des six nouvelles lettres

Aux vingt-deux lettres empruntées de l'alphabet nabatéen, les philologues arabes ont inventé six nouvelles lettres. Il s'agit des lettres suivantes : ث (U+062B), خ (U+06DE), ذ (U+0630), ض (U+0636), ظ (U+0638), غ (U+063A).

Ajout des points diacritiques

Étant donné que l'alphabet nabatéen, dont dérive l'alphabet arabe, n'était pas à l'origine muni de points diacritiques, il restait encore à régler le problème de différenciation des lettres qui partageaient une forme identique. Des petits points noirs ont été utilisé comme marques de différenciation. Les points étaient placés au-dessus et au-dessous du contour de la lettre, individuellement ou par groupes de deux ou trois.

Exemples : ب (U+0628), ت (U+062A), ث (U+062B) ;

ف (U+0641), ق (U+0642) ;

ح (U+062D), خ (U+062E), ج (U+062C), etc.

Vocalisation

Étant donné que l'écriture arabe est une écriture purement consonantique, pour préciser la prononciation, dix signes ont été inventés. Il s'agit de trois voyelles brèves et de sept signes orthographiques qui s'ajoutent au-dessus et au-dessous des consonnes.

Les trois voyelles brèves sont :

- **fatha** « ◌َ U+064E» (elle surmonte la consonne et se prononce comme un « a » en français) ;
- **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).

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 :

1. Dynastie califienne de 750 à 1258. la dynastie tire son nom de son ancêtre, l'oncle du prophète, Al-'Abbās banū 'Abd al-Muttalib banū Hašim.

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** « ٱ 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. Les consonnes arabes figurent dans le tableau n° 1.

1.3. *Extension*

Bien que l'écriture arabe aie vu le jour avant l'Islam, elle ne s'est étendue qu'avec la propagation de celui-ci. En effet, entre les années 633 et 645, les Arabes avaient organisé les premières grandes conquêtes de l'Islam dans trois directions : vers l'Asie Mineure et Constantinople, en Afrique du Nord et en Espagne et vers l'Asie Centrale et l'Inde. Dans tous les pays conquis, la langue et l'écriture arabes se sont imposées pour l'apprentissage et la pratique de la religion qui exclut toute autre langue que l'arabe. *« Il va sans dire d'ailleurs que les progrès de l'islamisme ont joué un rôle capital dans la diffusion de l'écriture arabe. Le Coran, qui contient la révélation divine, ne doit être ni traduit, ni transcrit dans une écriture étrangère ; l'initiation à la religion islamique implique une connaissance, au moins rudimentaire, de la langue et l'écriture arabe »* [4].

N°	Caractères arabes	Valeurs Unicode	Translittération en caractères latins[2]	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*

2. La translittération est l'opération qui consiste à représenter les caractères d'une écriture alphabétique ou syllabique par les caractères d'un alphabet de conversion. En principe, cette conversion doit se faire caractère par caractère. Dans ce tableau nous avons donné pour chaque caractère arabe son équivalent en caractère latin d'après la norme ISO de translittération des caractères arabes en caractères latins : ISO 233.

Associé à un processus fondamentalement religieux, à savoir l'islamisation, l'alphabet arabe a servi à noter des langues qui non seulement sont très différentes de l'arabe, mais n'appartiennent pas à la famille sémitique. Certaines langues ont emprunté la totalité des lettres arabes (28 consonnes plus la hamza) telles que le persan, le turc ottoman, l'ourdou, le pashto, etc. D'autres n'ont emprunté que quelques-unes telles que le kurde sorani qui n'a emprunté que 21 lettres. Pour rendre certains phonèmes inconnus à la langue arabe tels que le « p », le « v », le « g », le « tch », etc., il a fallu inventer des nouvelles lettres par l'adjonction de points et de signes diacritiques aux lettres arabes les plus proches par la prononciation. Le tableau suivant illustre quelques exemples des nouvelles lettres créées à partir des lettres arabes :

Lettres arabes de base	**Lettres créées**
ب (U+0628)	پ (persan U+067E)
ق (U+0642)	ڤ (berbère U+06A4)
ل (U+0644)	ڵ (kurde sorani U+06B5)

Tableau 2. *Lettres arabes et lettres créées*

Cependant, la notation des voyelles constitue la difficulté essentielle à laquelle se sont heurtés les introducteurs de l'alphabet arabe. En dépit d'un grand nombre de consonnes, l'écriture arabe n'utilise que trois signes pour noter les trois voyelles brèves arabes (voir plus loin la section 3.1.1 : Vocalisation). Ces trois signes s'ajoutent soit au-dessus soit au-dessous des consonnes. Par conséquent, un tel système, appliqué à des langues où les voyelles sont nombreuses, donne des résultats insatisfaisants. Face à ce handicap, certaines langues ont définitivement abandonné l'alphabet arabe en adoptant l'alphabet latin. On cite à titre d'exemple la langue turque qui depuis 1928 s'écrit en caractères latins ainsi que la langue kurde kurmanji qui après 1930 se notait en caractères latins et en caractères cyrilliques alors qu'elle s'écrivait en caractères arabes depuis le XVI^{e} siècle.

Par contre, d'autres langues ont gardé l'alphabet arabe tout en substituant les voyelles brèves arabes par des lettres. C'est le cas de la langue kurde sorani qui utilise un groupement des deux et des trois lettres pour noter ses huit voyelles. À l'instar des consonnes, les voyelles changent de graphie en fonction de leur position dans le mot (initiale, médiane, finale et isolée). À l'initiale d'un mot les voyelles sont écrites toujours à l'aide du ئـ (la lettre arabe ya' dans sa forme initiale ى support de

hamza ء) qui est utilisée uniquement dans cette position, ainsi que dans quelques rares emprunts à l'arabe. Les huit voyelles kurdes sorani sont les suivantes[3] :

	Description	I.	M.	F.	Is.
ئـا **(a)**	Représente une voyelle ouverte longue légèrement arrondie. On peut la rapprocher du français « pâte ».	ئـا	ـا	ـا	ئـا
ئـو **(o)**	Représente une voyelle orale arrondie longue. On peut la rapprocher du français « dôme ».	ئـو	ـو	ـو	ئـو
ئـوو **(û)**	Représente une voyelle arrondie d'aperture minimale. On peut la rapprocher du français « lourd ».	ئـوو	ـوو	ـوو	ئـوو
وی **(ö)**	Représente une diphtongue formée par l'ouverture progressive des lèvres. On peut la rapprocher du français « ouais ». Cette voyelle ne se réalise pas à l'initiale d'un mot.	-	ـویـ	ـوی	ئـوی
ئـه **(e)**	Représente une voyelle moyenne non arrondie brève. On peut la rapprocher du français « sel ».	ئـه	ـه	ـه	ئـه
ئـ **(i)**	Représente une voyelle d'aperture moyenne. Elle n'est écrite qu'à l'initiale de mots généralement empruntés à l'arabe. On peut la rapprocher du français « le ».	ئـ	-	-	ئـ
ئـی **(ê)**	Représente une voyelle fermée antérieure longue. On peut la rapprocher du français « été ».	ئـیـ	ـیـ	ـی	ئـی

Tableau 3. *Les voyelles kurdes sorani*

Légende : *I. : forme initiale, M. : forme médiane, F. : forme finale et Is. : forme isolée*

3. Il s'agit d'une prononciation française.

Concernant le classement alphabétique des lettres, l'étude que nous avons menée sur plusieurs alphabets nous a permis de constater que les alphabets en question adoptent le classement morphologique maghrébin[4] sauf pour les trois dernières lettres qui sont classées : و (U+0648) / ه (U+0647) / ي (U+0649) et non pas ه/و/ي. Par ailleurs, les lettres arabes modifiées, se positionnent juste après les lettres arabes à partir desquelles elles étaient créées.

En plus des lettres, les chiffres hindous utilisés par les pays arabes de l'Orient ont été également adoptés par plusieurs pays musulmans non arabes avec de légères modifications graphiques (voir plus loin la section 3.2.5).

2. Le traitement informatique de l'écriture arabe

2.1. *Problèmes posés par l'écriture arabe*

Contrairement à l'écriture latine qui possède depuis l'époque romaine deux styles fondamentaux : l'écriture manuscrite et les caractères utilisés par les graveurs pour les inscriptions lapidaires qui ont été utilisés par la suite pour confectionner les caractères mobiles de la typographie, l'écriture arabe est une et l'imprimerie ainsi que l'outil informatique, jusqu'à présent, se contente de reproduire fidèlement la structure manuscrite et cursive de l'écriture.

Parmi les caractéristiques qui distinguent ce style d'écriture arabe par rapport au style lapidaire d'écriture latine et qui, par conséquent, ont engendré des difficultés lors de son introduction aux nouvelles technologies, à savoir l'imprimerie d'abord et l'informatique ensuite, nous retiendrons les caractéristiques suivantes :

2.1.1. *La multitude de graphismes*

En s'écrivant, les lettres arabes se lient les unes aux autres. Cet usage entraîne quatre morphologies différentes d'une même lettre en fonction de son emplacement dans le mot : initiale, médiane, finale et isolée à l'exception des six lettres qui ne

4. Contrairement à l'alphabet latin qui a un système standard de classement des lettres allant de la lettre **a** à la lettre **z** sans tenir compte de la forme minuscule ou majuscule, l'écriture arabe possède trois systèmes de classements différents. Le premier système se base sur les valeurs numériques des lettres. En effet, autrefois les Arabes ont utilisé leurs lettres alphabétiques comme de véritable signes de numération en attribuant à chacune une valeur numérique. La succession de lettres correspond aux valeurs numériques respectives. Ce classement est actuellement en usage dans les pays arabes de l'Orient. Le second se base sur la morphologie des consonnes, c'est ainsi que les lettres ayant des formes graphiques ressemblantes se suivent. Ce système est actuellement en usage dans les pays du Maghreb. Le dernier type de classement se base sur la physiologie phonétique. Cet arrangement très variable, n'est utilisé que dans les études phonétiques.

possèdent que deux formes seulement : ا (U+0627), د (U+062F), ذ (U+0630), ر (U+0631), ز (U+0632), و (U+0648). Exemples :

Lettres arabes	Initiale	Médiane	Finale	Isolée
ب	بـ (U+FE91)	بـ (U+FE92)	ب (U+FE90)	ب (U+FE8F)
ص	صـ (U+FEBB)	صـ (U+FEBC)	ص (U+FEBE)	ص U+FEBD)
ع	عـ (U+FECB)	عـ (U+FECC)	ع (U+FECA)	ع (U+FEC9)
ز	ز (U+FEAF)	ز (U+FEB0)	ز/ز	ز (U+FEAF)

Tableau 4. *Les formes des lettres arabes*

2.1.2. *Les ligatures*

L'écriture arabe connaît trois types de ligatures : les ligatures contextuelles, les ligatures linguistiques et les ligatures esthétiques. Une ligature contextuelle est une chaîne de caractères prenant des formes spéciales suivant leur position dans le mot en obéissant à des règles grammaticales strictes et liées uniquement à l'écriture. Les ligatures linguistiques sont indispensables pour l'écriture d'une langue donnée et obéissant à des règles grammaticales. Souvent elles ont un statut de lettre et parfois même une place à part dans le dictionnaire, ce qui les rapproche des digraphes. Les ligatures esthétiques sont des graphies optionnelles qui existent pour des raisons esthétiques, de lisibilité et/ou de tradition. On peut les remplacer par leurs composantes sans changer la validité grammaticale, ou le sens du texte [1, 3].

En arabe, les ligatures contextuelles et linguistiques sont obligatoires et indispensables pour composer un texte manuscrit, typographique ou dactylographique. Comme ligature linguistique, l'arabe ne connaît que la ligature *lâm-alif*. Bien qu'il s'agisse de la lettre *lâm* (U+0644) sous forme initiale suivie de la lettre *alif* (U+0627) sous forme finale, il est impensable d'écrire " لـا " mais toujours لا (U+FEFB).

ا + ل = لا

Comme ligatures contextuelles on aura à titre d'exemple :

بـعث = ث +ع + ب

ب + ع + ث + ت = بـعثـت et ainsi de suite...

Contrairement aux ligatures contextuelles et linguistiques, les ligatures esthétiques sont optionnelles. Elles sont utilisées dans la paléographie et la typographie de qualité.

Exemples des ligatures esthétiques au début des mots :

ﲢ pour ح + م + ل

ﳾ pour م + ش

ﲝ pour ج + ن

Exemples des ligatures esthétiques à la fin des mots :

ﻳﺮ pour ي + ر

ﻧﻲ pour ي + ن

ﺗﻲ pour ي + ت

Exemples des ligatures esthétiques qui forment des mots en entier :

ﷻ : cette ligature esthétique est composée des deux mots جل جلاله

ﷺ : cette ligature est composée des quatre mots صلى الله عليه و سلم

« *La richesse de l'imagination des calligraphes a conduit à rendre très coûteuse l'imprimerie des textes arabes à l'aide des techniques traditionnelles. Les cases pouvaient contenir jusqu'à 1000 caractères* » [2]. En informatique, les ligatures rendent difficile la reconnaissance optique, l'affichage et l'impression des caractères.

2.1.3. *La vocalisation*

Pour préciser la prononciation du texte purement consonantique, les voyelles brèves et les signes orthographiques sont ajoutés au-dessus et au-dessous des consonnes. Chaque forme de consonne est susceptible de porter chacun des dix signes et souvent deux d'entre eux superposés. Du point de vue technique, les ligatures esthétiques rendent pratiquement impossible la position exacte de signes de vocalisation surtout lorsqu'il s'agit d'une superposition de trois ou quatre lettres. En plus, l'usage de la vocalisation réduit considérablement la rapidité de la composition typographique qui ne dépasse pas les 60 mots à la minute, alors qu'avec les machines à caractères latins on dépasse les 100 mots à la minute [8].

2.1.4. *La normalisation des caractères*

En s'écrivant, certaines lettres arabes montent très haut comme le *alif* surmonté du hamza (أ U+0623), le *kâf* (ك U+0643) et le *lâm* (ل U+0644), d'autres descendent très bas comme le *mîm* final (ﻢ U+FEE2), le *hâ* final (ﺦ U+FEA6) ainsi que le *gîm* final (ﺞ U+FE9E) et enfin on trouve des lettres très petites comme le *bâ* (ﺑ U+FE92), le *mîm* (ﻣ U+FEE4), et le *tâ* (ﺗ U+FE98) médianes. En typographie, pour que les petites lettres soient lisibles et distinguables, on est obligé de bien espacer les lignes. Par conséquent, on met beaucoup de mots dans une ligne mais beaucoup moins de lignes dans une page qu'avec les caractères latins. En plus l'usage des signes de vocalisation oblige d'espacer encore plus les lignes.

À ces problèmes s'ajoutent :

- l'écriture et la lecture de l'arabe s'effectuent de droite à gauche ;
- les formes minuscules et majuscules des lettres sont inexistantes ;
- les pays arabes font usage de deux types de graphies pour représenter les chiffres : les chiffres arabes et les chiffres hindous (voir section 3.2.5 : Chiffres).

2.2. *Tentatives de simplification*

Les problèmes posés par l'écriture arabe ont déclenché depuis les années trente une prise de conscience d'un certain nombre d'académies arabes, de la Ligue des États Arabes ainsi que de certains organismes culturels qui étaient convaincus que la question de l'écriture est un problème grave auquel il faut donner rapidement une solution.

En 1938 a été créée pour la première fois, au sein de l'Académie de Langue Arabe du Caire (Égypte), une « commission de la réforme des lettres arabes » qui a fixé comme objectif de simplifier l'écriture arabe.

En 1945, l'Académie de Langue Arabe du Caire a organisé un concours du meilleur projet de réforme de l'écriture arabe (avec récompense de 1000 livres égyptiennes à l'appui). La majorité des pays arabes et quelques pays musulmans ainsi que des pays européens ont participé à ces projets. En 1968, l'Académie a constitué une commission pour étudier les projets reçus dont le nombre dépasse les 200 projets de réforme. L'ensemble des projets est reparti en trois grandes catégories :

- ceux qui gardent les lettres arabes et les signes de vocalisation (signes orthographiques et voyelles brèves) ;
- ceux qui gardent les lettres arabes et remplacent les signes de vocalisation par d'autres signes incorporés aux mots ;
- ceux qui emploient les caractères latins [8].

Malgré cette prise de conscience des problèmes engendrés par l'écriture arabe, aucun projet de réforme n'a été retenu. L'échec des tentatives de réforme et surtout de la normalisation de l'alphabet arabe peut être à l'origine de plusieurs raisons dont les plus importantes sont :

Raisons religieuses

Le Coran, livre sacré, occupe une place centrale dans la société arabo-musulmane. Le Coran a été révélé en langue arabe. De ce fait, la langue arabe, pour les Arabes, non seulement est un instrument de communication comme les autres langues du monde, mais aussi elle est sacrée. C'est la langue que Dieu a choisie pour révéler son message au prophète Mahomet. Par conséquent, changer l'écriture est interprété comme un changement du Coran et une coupure de l'Islam, ce qui explique la farouche opposition qui a été provoquée dans l'ensemble des pays arabes qui ont estimé qu'au bout d'une génération ceux qui apprendront la nouvelle graphie seront incapables de lire la parole de Dieu.

Raisons culturelles

Les Arabes sont très attachés à leur langue et à leur écriture. Les adversaires des réformes ont considéré tout changement de l'écriture comme une condamnation d'un des aspects les plus importants de la civilisation arabe et une rupture radicale entre le présent et la passé d'une part, ainsi que tous les trésors de la littérature arabe et l'héritage des ancêtres, d'autre part. Cette fierté de parler l'arabe et d'écrire en arabe classique s'aperçoit clairement dans les discours de certains opposants de réforme.

Raisons nationalistes

Face à l'occident, les Arabes veulent affirmer leur identité nationale en se référant à la richesse de leur civilisation, de leur culture et surtout de leur langue. Accepter de réformer l'écriture arabe en adoptant l'alphabet latin ou même d'incorporer les voyelles aux consonnes à l'instar de l'alphabet latin est considéré comme une forme de dépendance vis-à-vis de l'occident et une reconnaissance de l'inadéquation de l'arabe, langue du Coran et fierté des Arabes.

Il est bien évident que l'échec des tentatives de simplification de l'écriture arabe avait engendré des problèmes lors de traitement de l'alphabet arabe avec cette nouvelle technologie.

2.3. *Les normes de codage des caractères arabes*

Le développement de l'informatique aux États-Unis dans les années 1960 a donné lieu à la création d'une norme de codage qui va marquer le domaine jusqu'à nos jours. Il s'agit de la norme ASCII (ISO/CEI 646)[5]. Pour permettre le codage d'autres systèmes d'écriture, 12 codes du jeu graphique de l'ISO/CEI 646 sont réservés aux besoins nationaux. Avec seulement douze codes, l'ISO/CEI 646 ne permet pas de représenter l'alphabet arabe complet : consonnes, voyelles brèves et signes orthographiques. Or, il est impossible de traiter efficacement les données lexicographiques arabes sans les signes de vocalisation. Par conséquent, afin de

5. Voir article de Jacques André dans ce numéro.

conquérir le marché informatique assez important dans les pays arabes, certains constructeurs informatiques ont cherché à définir des jeux codés pour l'alphabet arabe. Malheureusement, la hâte et la concurrence des solutions industrielles ont conduit à la définition de jeux incohérents.

Jusqu'en 1976, il n'existait aucun code normalisé de transmission de données en caractères arabes qui comportait l'alphabet arabe complet, signes de vocalisation inclus. L'établissement d'un code arabe unifié était une affaire urgente et c'est en 1975 que fut pour la première fois dans le monde arabe posé le problème d'établir un code unifié pour la transmission de données en caractères arabes équivalant à la norme ISO/CEI 646. Dès lors, les réunions regroupant constructeurs et représentants arabes se multiplièrent.

En 1982, le comité technique n°8 de l'ASMO[6] (Organisation Arabe de Normalisation et de Métrologie) a développé la norme ASMO 449. C'est une norme de codage de caractères sur sept bits prévue pour être utilisée pour l'échange d'informations ainsi que pour des applications de traitement de données et de textes en langue arabe. Elle a été adoptée par la suite par l'ISO comme une norme internationale ISO 9036 [5]. Trois ans après et dans le but de mettre en place un tableau de code arabe indépendant qui permettra la conception d'un ordinateur qui fonctionne uniquement en arabe, le même comité technique a développé la norme ASMO 662, une norme de codage sur 8 bits (256 caractères avec possibilité de coder les caractères des langues qui utilisent les caractères arabes).

En 1988, le comité technique n°8 de l'ASMO a développé la norme ASMO 708 pour gérer un bilinguisme arabe/latin. C'est un jeu de 146 caractères graphiques identifiés portant le nom alphabet latin/arabe et la représentation codée de chacun de ces caractères sur un octet. La norme attribue à la zone gauche des caractères graphiques le jeu de caractères de la norme ISO/CEI 646 et à la zone droite le jeu de caractères arabes défini dans la norme ASMO 449. La norme ASMO 708 a été adoptée par l'ECMA comme une norme internationale sous la référence ECMA-114 et a été également adoptée par l'ISO comme une norme internationale dans la série des normes ISO 8859-n sous le nom ISO 8859-6 [6].

Outre ces trois jeux de caractères, plusieurs autres jeux ont été définis et qui sont actuellement disponibles. On cite à titre d'exemple :

- MS Windows Arabic Code Page 1256[7] ;
- Arabic Mac Code Page[8] ;
- Arabic Windows 3X Code Page[9] ;

6. ASMO a été dissoute en 1990.
7. http://www.microsoft.com/globaldev/reference/sbcs/1256.htm
8. ftp://ftp.unicode.org/Public/MAPPINGS/VENDORS/APPLE/ARABIC.TXT
9. http://members.aol.com/ArabicLexicons/codepages/codepage_win3x.htm

- Code Page 864 Dos Arabic[10], etc.

Les codes correspondant à des caractères ASCII (les 128 premiers caractères) sont les mêmes pour tous ces jeux de caractères. Cependant, il n'en va pas de même pour les caractères arabes, d'où une perte plus ou moins importante des données quand on passe d'un jeu de caractères à un autre. Par exemple, le mot arabe suivant « قوي » (fort en français) est codé en ISO 8859-6 sous la forme de la séquence des codes numériques suivante : 226 (ق), 232 (و) et 234 (ي). Si on passe de l'ISO 8859-6 à la page de code Windows 1256, la même séquence des codes numériques donnera à l'affichage la suite des caractères suivante : êèâ. Cette transformation s'explique par le fait que les deux jeux de caractères attribuent des codes différents pour les trois lettres arabes de notre exemple. Ce problème se rencontre notamment lorsqu'on essaye d'ouvrir un site Web codé en ISO 8859-6 en utilisant la page de code Windows 1256.

Outre l'écriture arabe de base, certaines pages de codes codent aussi certains caractères utilisés par des langues qui ont adopté l'alphabet arabe. Le tableau suivant identifie quelques-uns de ces caractères codés dans les pages de codes CP1256 et CP Arabic Mac :

Caractères	**CP 1256**		**CP Arabic Mac**	
	V. hex.	V. num.	V. hex.	V. num.
پ	81	129	F3	243
چ	8D	140	F5	245
ژ	8E	142	FE	254
گ	90	144	F8	248

Tableau 5. *Les pages de codes CP 1256 et CP Arabic Mac*

On peut remarquer sur ce tableau que même pour ces caractères le problème de compatibilité entre les jeux de caractères persiste encore. Le même problème est observé également avec les caractères latins accentués. En effet, si on passe de l'ISO Latin1 à la page de codes Macintosh ou à la page de codes MS-DOS 850, les codes numériques de ces caractères changent, ce qui provoque une perte des données.

10. http://www.kostis.net/charsets/cp864.htm

3- Le standard Unicode et la norme ISO/CEI 10646

La difficulté de gérer des applications multilingues, dans un contexte où existent un grand nombre de normes différentes d'un pays à l'autre et parfois au sein d'un même pays, a conduit à l'idée d'un jeu de caractères dit universel, dans lequel la plupart des caractères permettant de transcrire les langues existantes seront codés. De cette idée sont nés la norme ISO/CEI 10646 et le standard Unicode[11].

À la suite de négociations officielles sur une convergence des deux codes, les deux répertoires ont été fusionnés en 1992. 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/CEI 10646 mais qui sont absents du standard Unicode. Par rapport à la version 3.0 qui code 49 194 caractères au niveau du BMP, la version 3.1 ajoute 44 946 nouveaux caractères pour atteindre un total de 94 140 caractères. Elle constitue, en effet, la première version du standard Unicode qui code des caractères en dehors de l'espace de codage du PBM. Elle est rigoureusement identique aux deux normes ISO : ISO/CEI 10646-1 : 2000 (Part 1 : Architecture and Basic Multilingual Plane) et ISO/CEI 10646-2 : 2000 (Part 2 : Supplementary Planes).

3.1. *Principes généraux du traitement de l'écriture arabe*

À la différence de l'écriture latine, l'écriture arabe ainsi que les écritures qui sont dérivées de l'arabe de base se caractérisent notamment par sa directionnalité droite-gauche, par sa nature cursive et par ses signes de vocalisation qui s'ajoutent au-dessous et au-dessus des caractères. Ces trois caractéristiques constituent en fait les problèmes majeurs qu'elle pose face aux technologies informatiques. Pour remédier à ces problèmes, le standard Unicode offre toute une panoplie de codes de formatage et des algorithmes permettant par conséquent un traitement informatique fiable de l'écriture arabe et des écritures qui en dérivent.

3.1.1. *Bidirectionnalité*

Les textes Unicode, quelle que soit l'écriture, sont stockés en mémoire en ordre logique. Cet ordre correspond à l'ordre dans lequel le texte est saisi au clavier. Parfois, l'ordre des caractères à l'affichage ou à l'impression diffère de l'ordre logique.

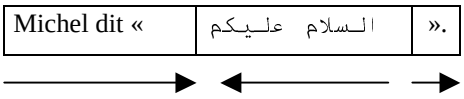
11. voir article de Patrick Andries dans ce numéro.

Exemple d'un texte bilingue (français arabe) :

ordre de saisie :

M	i	c	h	E	l		d	i	t		«		ا	ل	س	ل	ا	م		ع	ل	ي	ك	م	»	.

ordre à l'affichage :



À l'affichage, le glyphe qui représente le premier caractère du texte français apparaît à gauche. Cependant, le premier caractère logique du texte arabe est représenté par le glyphe arabe le plus proche de la marge droite. Les glyphes arabes suivants sont disposés vers la gauche. Le standard Unicode décrit l'ordre logique que les caractères doivent respecter en mémoire. Lorsqu'il s'agit d'un texte qui possède la même direction horizontale (gauche à droite, droite à gauche), l'ordre d'affichage du texte se fait sans ambiguïté. Cependant, l'ordre d'affichage des caractères peut parfois être ambigu lorsqu'il s'agit d'un texte bidirectionnel (un mélange de texte allant de gauche à droite et de droite à gauche).

Dans le standard Unicode, tous les caractères possèdent un type directionnel normatif. Le type bidirectionnel de caractère est une valeur attribuée à chaque caractère Unicode, y compris les caractères non affectés. Les types directionnels gauche-à-droite et droite-à-gauche sont qualifiés de forts, on appelle les caractères dotés de ces types *des caractères à forte directionnalité*. Les types gauche-à-droite comprennent la plupart des caractères alphabétiques et syllabiques ainsi que les caractères idéographiques Hans. Les types droite-à-gauche comprennent l'arabe, l'hébreu, le syriaque, le thâna ainsi que la majorité de la ponctuation associée à ces écritures. Les types directionnels associés aux chiffres sont qualifiés de type faible, ces types de caractères sont dits à *faible directionnalité*. Par conséquent, l'ordre d'affichage du texte bidirectionnel dépend des propriétés directionnelles des caractères qui composent le texte qui sont généralement suffisantes pour que le texte soit affiché correctement.

Cependant, dans des circonstances particulières les propriétés directionnelles implicites des caractères ne permettent pas un affichage correct des caractères notamment lorsqu'il s'agit d'un texte composé de plusieurs systèmes d'écritures avec des directionnalités différentes (par exemple un texte composé d'un mélange de plusieurs langues : de l'arabe, du français et de l'hébreu). Afin de résoudre ce problème, le standard Unicode comprend un certain nombre de caractères de formatage directionnel explicites. Ces codes permettent de contrôler de manière précise l'ordre d'affichage des caractères. Le tableau suivant énumère ces codes :

N° CODE (HEX.)	NOM	ABRÉVIATION	SIGNIFICATIONS
U+200E	Marque gauche-à-droite	MGAD (LRM)	L'effet de ces deux codes est identique à celui d'un caractère à forte directionnalité. Ils sont conçus pour dissiper les incertitudes quant à la directionnalité de certains passages dans les textes bidirectionnels
U+200F	Marque droite-à-gauche	MDAG (RLM)	
U+202A	Enchâssement gauche-à-droite	EGAD (LRE)	Ce code indique qu'un passage de texte est enchâssé et doté d'une directionnalité droite-à-gauche.
U+202B	Enchâssement droite-à-gauche	EDAG (RLE)	Ce code indique qu'un passage de texte est enchâssé et doté d'une directionnalité gauche-à-droite.
U+202D	Forçage gauche-à-droite	FGAD (LRO)	Ce code permet de forcer la directionnalité des caractères gauche-à-droite. Avec ce code le mot arabe suivant : ابـحـرم s'affiche مـرحبـا
U+202E	Forcâge droite-à-guche	FDAG (RLO)	Ce code permet de forcer la directionnalité des caractères droite-à-gauche. Avec ce code le mot français suivant : Salut s'affiche tulaS
U+202C	Dépilement de formatage directionnel	DFD (PDF)	Ce code permet de rétablir l'état bidirectionnel à ce qu'il était avant le dernier FDAG, FGAD, EDAG ou EGAD.

Tableau 6. *Les codes de formatage directionnels*

Tous ces codes ne s'appliquent qu'au paragraphe courant et leur effet s'annule au prochain séparateur de paragraphe U+2029.

3.1.2. *Nature cursive de l'écriture arabe*

Même sous sa forme imprimée, l'arabe est fondamentalement une écriture cursive : les différents caractères formant le mot sont liés entre eux. Par conséquent, Le même caractère peut avoir quatre formes différentes en fonction de sa position dans le mot (initiale, médiane, finale et isolée). Exemples :

- ب (Bâ,U+0628) + ح (Hâ, U+062C) = بـج (Bâ forme initiale et Hâ forme finale).

- ب (Bâ,U+0628) + ح (Hâ, U+062C) + ر (Râ, U+0631) = بـحر (Bâ forme initiale, Hâ forme médiane et Râ forme finale), etc. Ces liaisons cursives sont mises en œuvre dans la quasi-totalité des polices arabes.

Cependant, certaines écritures qui utilisent l'alphabet arabe, interdisent dans certains cas la liaison entre les lettres alors qu'elles se côtoient sans un espace de séparation. C'est le cas du suffixe pluriel perse, certains noms de famille perses et les voyelles en turc ottoman. Afin de remédier à ce problème, le standard Unicode fournit aux utilisateurs deux codes de formatage **U+200C** (ALSC : Antiliant sans chasse, en anglais ZWNJ : ZERO WIDTH NON-JOINER) et **U+200D** (LSC : Liant sans chasse, en anglais ZWJ : ZERO WIDTH JOINER). L'Antiliant sans chasse (U+200C) a pour but de dissocier les ligatures et d'empêcher la formation de liaisons cursives entre deux caractères et le liant sans chasse (U+200D) a pour but d'afficher, si possible, les caractères adjacents sous une forme plus liée qu'il ne l'aurait été en son absence. Ces caractères ne modifient pas le processus de sélection contextuelle, ils changent plutôt le contexte d'un caractère particulier. À l'affichage, l'effet de ces caractères dépend du contexte dans lequel ils apparaissent. C'est ainsi que si on ajoute un liant sans chasse entre deux caractères qui sont déjà liés de manière cursive il n'aura aucun effet (le cas de notre exemple précédemment cité). De même si on ajoute un antiliant sans chasse entre deux caractères qui ne sont pas reliés il ne produira aucun effet (exemple un texte français).

Dans la version 3.1 du standard Unicode, la sémantique de ces deux codes de formatage a été beaucoup étendue. L'usage du code ALSC entre deux caractères empêche à la fois la liaison cursive et la formation de ligature entre ceux-ci lors du rendu. Par contre, l'usage du code LSC permet la liaison cursive ainsi que la formation de ligatures entre eux (si c'est possible). Exemples :

- ل (U+0644) + ALSC (U+200C) + ج (U+062C) = لج (le ALSC a empêché la liaison cursive et la formation de ligature entre ces deux caractères. Habituellement, ces deux caractères s'affichent لـج) ;
- ل (U+0644) + LSC (U+200C) + ج (U+062C) = ﺟ (le LSC a permis la liaison cursive et la formation d'une ligature entre ces deux caractères. Sans le LSC, ces deux caractères s'affichent لـج) ;

Dans certains cas, les deux codes de formatage LSC et ALSC peuvent être combinés. Par exemple, si on veut que les deux caractères de notre exemple ل et ج

soient liés sans formation d'une ligature entre eux, la séquence suivante peut être utilisée :

- ل (U+0644) + LSC (U+200D) + ALSC (U+200C) + LSC (U+200D) + ج (U+062C) = لـج (لـ forme initiale et ج forme finale).

De même, si on veut que la lettre ل prenne une forme cursive et que la lettre ج prenne la forme isolée, la séquence suivante peut alors être utilisée :

- ل (U+0644) + LSC (U+200D) + ALSC (U+200C) + ج (U+062C) = لـج (لـ forme initiale et ج forme isolée).

Le tableau suivant illustre quelques exemples de l'utilisation de ces deux codes :

Formes habituelles	**ALSC**	**LSC ALSC LSC**	**LSC ALSC**	**LSC**
Fi ou fi	Fi	Fi	Fi	fi
لا	ل ا	لـا	لـ ا	لا
وج	وج	وج	وج	وج

Tableau 7. *Les codes ALSC et LSC*

3.1.3. *Signes de vocalisation*

Dans le standard Unicode, les signes de vocalisation (voyelles brèves et signes orthographiques) sont dits « caractères combinatoires ». Ces caractères sont destinés à s'afficher par rapport à un caractère de base. Par définition, un caractère de base est un caractère qui ne se combine pas graphiquement avec des caractères qui le précèdent et qui n'est ni un caractère de commande ni de formatage. Cette absence de combinaison graphique de la part des caractères de base ne leur interdit pas d'adopter des formes contextuelles ou de prendre part à des ligatures. Par contre, un caractère combinatoire est un caractère qui se combine avec un caractère de base qui le précède. Comme caractères combinatoires on cite les accents, les points hébreux, les voyelles brèves et les signes orthographiques arabes, les matras indiennes et autres diacritiques. Ils sont représentés dans le tableau de codes sous forme de caractères au-dessus, au-dessous ou à travers un cercle en pointillés. Ce cercle pointillé représente le caractère de base en question. La position du caractère combinatoire par rapport à ce cercle pointillé indique la position qu'il doit prendre par rapport au caractère de base. On retrouve dans les différents blocs du standard Unicode des caractères combinatoires utilisés principalement avec les écritures représentées par ces blocs. Cependant, l'affectation d'un caractère combinatoire à un bloc de caractères n'empêche pas qu'il soit combiné avec des caractères de base

codés dans d'autres blocs. En effet, dans le standard Unicode, toutes les suites de caractères sont permises.

Dans le standard Unicode, tous les caractères combinatoires doivent suivre le caractère de base auquel ils s'adjoignent. Cette logique d'ordonnancement correspond à la logique de l'écriture arabe et indienne dont les signes diacritiques suivent toujours les caractères de base auxquels ils se placent. Ainsi, la suite de caractères Unicode U+0061 a (Lettre minuscule latine a) + U+0308 ¨ (Diacritique tréma) + U+0075 u (Lettre minuscule latine u) code sans équivoque « äu » et non « aü ».

Pour l'écriture arabe et ses dérivées, le standard Unicode code 33 caractères combinatoires au niveau de la rangé 6. La figure suivante illustre une partie du bloc de caractères arabes de base et étendus (rangée six) avec quelques-uns des caractères combinatoires qui figurent au niveau des colonnes six et cinq et un seul à la colonne 8.

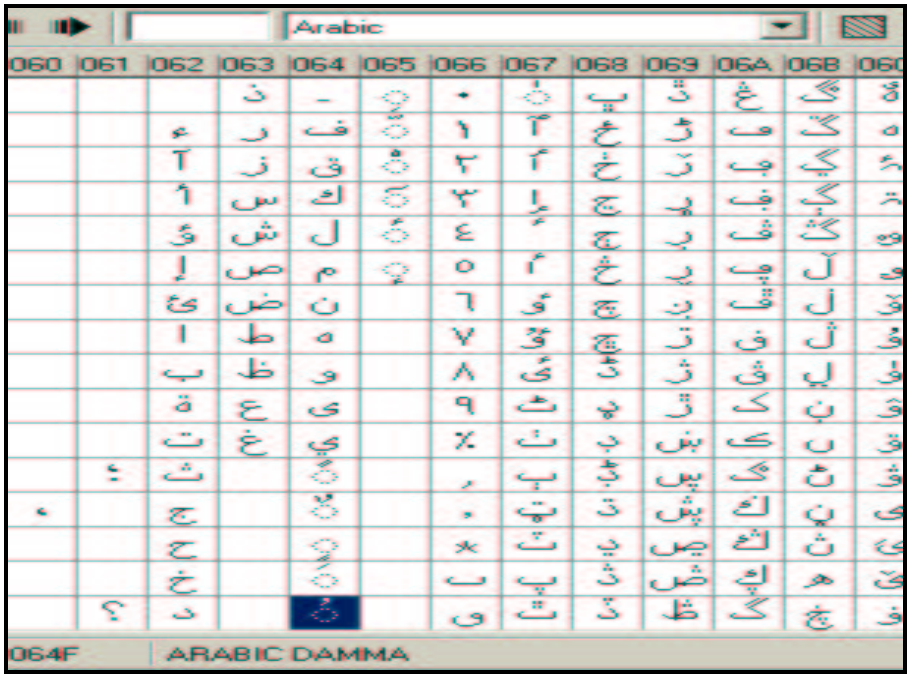


Figure 1. *Les caractères combinatoires arabes*

Outre les signes de vocalisation (les trois voyelles brèves et les sept signes orthographiques), on compte un grand nombre de signes qui sont utilisés

principalement dans les textes coraniques comme à titre d'exemple les deux caractères 06DD ۝ (FIN DE AYAH ARABE, ARABIC END OF AYAH) et 06DE ۞ (DÉBUT DE ROUB EL HISB ARABE, ARABIC START OF RUB EL HISB)[12].

En arabe de base, dans un texte vocalisé, on fait usage de 10 signes de vocalisation qui s'ajoutent au-dessous ou au-dessus des consonnes. Seules les trois voyelles brèves et les trois signes de tanwin peuvent parfois être combinés avec le signe orthographique « chadda » alors que les restes des signes orthographiques sont toujours notés séparément. D'après la logique d'Unicode, une lettre arabe voyellisée sera codée de la manière suivante : caractère consonne de base + signe orthographique + caractère voyelle.

3.2. *Le codage des caractères arabes dans le PBM*

Le standard Unicode et la norme ISO/CEI 10646 réserve 942 codes pour coder l'écriture arabe ainsi que les 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 (PMB).

3.2.1. *Rangée 6*

La rangée 6 est destinée à coder les caractères arabes de base et étendus. Les caractères arabes de base (positions U+0600 à U+0652) sont au nombre de 48. Ils sont identiques au jeu graphique de la norme ASMO 449/ISO 9036 et ASMO 708/ISO 8859-6 (Partie droite). Ceci est d' ordre à assurer une compatibilité avec les anciennes normes arabes de codage.

Le jeu étendu des caractères arabes couvre les caractères des alphabets qui sont dérivés de l' alphabet arabe de base comme l' alphabet persan, urdu, pashto, kurde

12. Le Coran se compose de 114 sourates ou chapitres. Chaque sourate est composée d'un certain nombre des versets (en arabe Ayah آية). La plus longue sourate comprend 286 versets et la plus courte comprend seulement 3 versets. Le nombre total des versets dans le Coran est de 6247. Dans chaque sourate, les versets sont numérotés en utilisant des chiffres entourés par un symbole ornemental utilisé comme motif de décoration. Aux fins de récitation, le Coran se divise en 30 parties d'égale longueur. Chacune de ces parties s'appelle en arabe juz (جزء). Pour la même raison, l'Afrique du Nord divise l'ensemble du Coran en soixante sections (en arabe Hisb حزب) et chaque Hisb se divise en quatre sous-section (en arabe en arabe Roub el Hisb ربع الحزب). À l'instar des versets, on fait usage d'un symbole ornemental pour indiquer le numéro du Hisb et sa sous-section. Le même symbole est aussi utilisé pour indiquer le numéro de la partie (le juz'). Le caractère combinatoire 06DD représente le symbole ornemental qui englobe le numéro de chaque verset et le caractère combinatoire 06DE représente le symbole ornemental qui englobe le numéro du Hisb ainsi que le numéro de la partie. La taille de chiffres utilisés dans la numérotation est réduite au rendu.

sorani, etc. Outre les caractères des alphabets, la rangée 6 contient aussi des symboles religieux musulmans comme, à titre d'exemple, le symbole qui indique la fin d'un verset dans le Coran (U+06DD), etc.

Respectant le principe qu'un caractère même s'il est utilisé par plusieurs écritures ne se voit attribuer qu'un seul code dans le tableau de code, la majorité des signes de ponctuation ne sont pas codés. Seuls les signes qui sont propre à l'écriture arabe comme la virgule arabe (، : U+060C), le point-virgule arabe (؛ : U+061B), le point d'interrogation arabe (؟ : U+061F) et le symbole pour cent arabe oriental (٪ : U+066A) sont codés. La rangée 6 code 208 caractères.

3.2.2. *Rangée FB*

La rangée FB (positions U+FB50 à U+FBFF) est destinée à coder les variantes contextuelles des lettres arabes étendues (isolées, initiales, médianes et finales). Elle code 143 formes.

3.2.3. *Rangées FC et FD*

Dans l'écriture arabe ainsi que dans les écritures dérivées de l'arabe de base, il existe trois formes de ligatures : les ligatures contextuelles, les ligatures linguistiques et les ligatures esthétiques. Les rangées FC (positions U+FC00 à U+FCFF) et FD (positions U+FD00 à U+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, soit au total 451 formes y compris les 9 formes de combinaison de voyelles avec les signes orthographiques et deux symboles ornementaux qui jouent le rôle des parenthèses d' ouverture et de fermeture (U+FD3F Parenthèse droite ornée et U+FD3E : Parenthèse gauche ornée).

3.2.4. *Rangée FE*

Une partie de la rangée FE (positions U+FE70 à U+FEFF) est destinée à coder les variantes contextuelles de lettres arabes de base et de la ligature linguistique Lam-Alif (ﻻ). 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. Ces caractères sont codés pour des raisons de compatibilité avec des normes ou des systèmes préexistants qui utilisent ces formes en tant que caractères. Cependant, ils peuvent être remplacés par les caractères de la rangée six (U+0600-U+06FF).

3.2.5. *Chiffres*

Le standard Unicode et la norme ISO/CEI 10646 distinguent trois formes pour noter les chiffres utilisés dans l'écriture arabe de base ainsi que les écritures dérivées

de l'arabe de base. Le tableau suivant présente ces trois formes en donnant pour chacune le nom accordé par le standard et la norme, l'emplacement dans le tableau de code et enfin la représentation graphique des chiffres.

Nom	Numéros des caractères (hex.)	Formes
Européens (les chiffres arabes, utilisés par les pays arabes du Maghreb)	U+0030…U+0039	0123456789
Arabo-hindî (utilisés par les pays arabes de l'Orient ainsi que la Lybie)	U+0660…U+0669	٠١٢٣٤٥٦٧٨٩
Arabo-hindî orientaux (utilisés en Iran et au Pakistan)	U+06F0…U+06F9	۰۱۲۳۴۵۶۷۸۹

Tableau 8. *Les chiffres arabes*

Toutefois, on remarque que le standard Unicode et la norme ISO/CEI 10646 ne prévoient pas des codes pour coder les chiffres ourdous utilisés au Pakistan qui eux aussi ont été développés à partir des chiffres hindous.

3.3. *Produits Unicode supportant l'écriture arabe*

Tirant profit de leur importance économique et technologique, le standard Unicode et la norme ISO/CEI 10646 ont été vite adoptés par un grand nombre des développeurs d'applications et des constructeurs de matériels informatiques. Actuellement, ils sont exigés par de nombreux standards récents en provenance du W3C et de l'IETF tels que XML, HTML, XHTML, XSL, etc. L'IETF exige que tout nouveau protocole doit être en mesure d'utiliser le codage UTF-8 et que les protocoles existants qui utilisent d'autres jeux de caractères ou même qui utilisent un jeu de caractères par défaut que l'UTF-8, doivent supporter le codage UTF-8. En langage HTML, il suffit d'ajouter dans la section <head> d'un document une méta-information indiquant cela :

<meta http-equiv="content-type" content="text/html"; charset="UTF-8">

Avec le langage XML, la mention du codage UTF-8 doit être indiquée dans le prologue du document XML : **<?xml version="1.0" encoding="UTF-8"?>**

Dans le corps du document HTML ou XML, on peut insérer les caractères par référence aux numéros qu'ils ont dans le tableau des codes. Les références aux caractères peuvent être fournies soit en valeur hexadécimale ou en valeur décimale du caractère dans la table. Dans le premier cas, la référence est préfixée par &#x

suivis par le numéro hexadécimale du caractère. Dans le second cas, la référence est préfixée par &#, suivis par la valeur numérique du caractère dans la table suivi enfin par un point virgule.

Le codage Unicode est actuellement supporté par plusieurs systèmes d'exploitation (Windows NT et Windows 2000[13], Palm OS 4.0[14], MacOs 9.0[15], IBM AIX[16], Compaq's Tru64 UNIX[17], etc.) ainsi que par plusieurs langages de programmation (Java, JavaScript, IBM APL2, Microsoft VJ++, Visual Studio 7.0, Visual Basic, Perl 5.005, SoftQuad Xmetal 2.1, V1.5 pour Windows NT/2000 et V1.1 pour Windows 3.1, etc.).

Parmi la longue liste de produits qui supportent Unicode et par conséquent, supportent l'écriture arabe et les écritures qui en dérivent, on cite :

Polices de caractères arabes

- Arial Unicode MS : cette fonte contient pratiquement tous les caractères de la version 2 d'Unicode (51 180 glyphes dans la version 0.86). Elle couvre un grand nombre des caractères des rangées 6, FB, FC et FD. Elle est fournie avec Office 2000 Premium[18] ;
- Bitstream CyberBase (1249 glyphes dans la version beta v1.0)[19] ;
- Bitstream CyberBit (29,934 glyphes dans la version beta v2.0)[20] ;
- Code2000 (33,993 glyphes dans la version 1.10)[21] ;
- Courier New (1318 glyphes dans la version 2.82)[22] ;
- Microsoft Sans Serif (1090 glyphes dans la version 1.02). Elle est fournie avec Windows 2000 et Windows XP ;
- Tahoma (1227 glyphes dans la version 2.60). Elle est fournie avec Windows 2000 et Windows XP ;
- TITUS Cyberbit Basic (9568 glyphes dans la version 2.1)[23] ;
- Naqsh (1815 glyphes dans la version 2.0)[24] ;

13. http://www.microsoft.com
14. Palm OS 4.0 http://www.palmos.com/dev/tech/docs/palmos40/
15.Technical Note TN 1176 : Mac OS 9 http://developer.apple.com/technotes/tn/tn1176.html
16. AIX Version 4.3 http://www.rs6000.ibm.com/software/OS/aix43_specs.html#dev
17. http://www.tru64unix.compaq.com
18. http://office.microsoft.com/downloads/2000/aruniupd.aspx
19. ftp://ftp.netscape.com/pub/communicator/extras/fonts/windows/
20. ftp://ftp.netscape.com/pub/communicator/extras/fonts/windows/
21. http://home.att.net/~jameskass/
22. http://www.microsoft.com/truetype/fontpack/win.htm
23. http://titus.fkidg1.uni-frankfurt.de/unicode/tituut.asp
24. http://zaban.net/free_fonts

- Nesf (228 glyphes dans la version Oriental Newspaper 1/2/2000)[25] ;
- Traditional Arabic (530 glyphes dans la version 1.01). Elle est fournie avec Windows 2000 et Windows XP ;
- DecoType Naskh (256 glyphes dans la version 3.5d5e2 ;
- Arabic Newspaper (600 glyphes)[26];
- ClearlyU Arabic[27].

Éditeurs de textes multilingues

- Microsoft Word 2000 : il supporte le codage Unicode. Les caractères Unicode sont insérés sous forme des caractères spéciaux (Insertion/Caractères spéciaux). Word 2000 peut également être utilisé comme un éditeur HTML. Les caractères Unicode seront alors enregistrés sous leurs valeurs numériques préfixées par &# et suivies par un point virgule (voir figure n° 2);
- AbiWord : il tourne sous Windows et Unix. Il permet de lire et de sauvegarder des fichiers en formats UTF-8, Text, RTF, XHTML, LaTeX et Word [28] ;
- EmEditor : il tourne sous Windows 95, Windows 98, Windows Me, Windows NT4 et Windows 2000. Il permet de lire et de sauvegarder des fichiers en formats UTF-16, UTF-8, UTF-7 et en beaucoup d'autres jeux de caractères[29] ;
- UniRed : est un éditeur Unicode gratuit. Il tourne sous Windows 95, Windows 98, Windows Me, Windows NT4 et Windows 2000. Il permet de créer et de lire des fichiers en formats UTF-16, UTF-8 et en plusieurs autres jeux de caractères (ISO, Windows, Mac, etc.). Il est aussi un éditeur HTML ;
- UniPad : éditeur de textes qui tourne sous Windows 95, Windows 98, Windows Me, Windows 2000, Windows XP et Windows NT. Il permet de créer des fichiers en format UTF-8 et de créer des pages HTML multilingues[30] ;
- FrontPage 2000 et 2002 ;
- Muwse : éditeur HTML multilingue pour Macintosh. Il sauvegarde les fichier en format UTF-8[31] ;

25. http://digilib.sharif.ac.ir/fonts.php
26. http://crl.nmsu.edu/~mleisher/arabic24.html
27. http://crl.nmsu.edu/~mleisher/download.html
28. http://www.abisource.com/download/
29. http://www.emurasoft.com/emeditor3/
30. http://www.unipad.org/main/
31. http://www.makienterprise.com/muwse/muwse.html

- Pepper : éditeur HTML multilingue qui tourne sous Mac OS 9 et Mac OS X 10. Il permet d'importer et d'exporter des fichiers en formats UTF-8, UTF-16, ANSI, MacRoman, ShiftJIS, Big5 et toute la série des normes ISO 8859[32].

Navigateurs supportant Unicode et l'écriture arabe

- Internet Explorer 5, 5.5 et 6 (voir figure n° 3) ;
- Netscape 6 ;
- Tango[33] ;
- ICab Oreview 2.7[34] pour Macintosh.

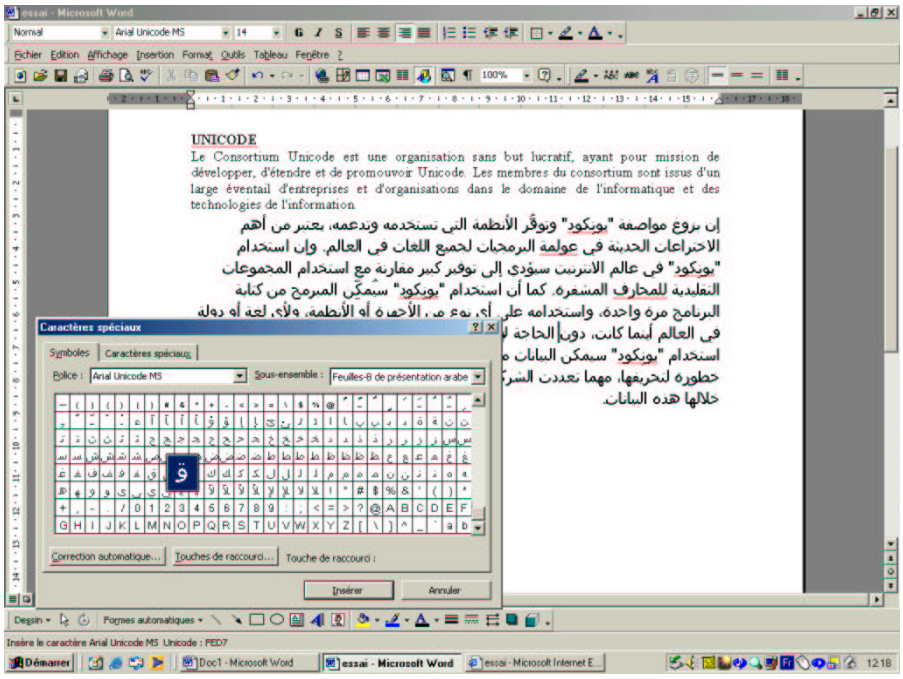


Figure 2. *texte bilinge arabe/français sous Word 2000 (version latine)*

32. http://www.hekkelman.com/pepper.html
33. http://www.alis.com
34. http://www.icab.de/download.html

Figure 3. *Les jeux de caractères supportés par le navigateur Internet Explorer6*

4. CONCLUSION

Suite à l'échec de toutes les tentatives de simplification et de réformes de l'écriture arabe en vue de son adaptation aux exigences de la vie moderne notamment aux nouvelles technologies, les mêmes problèmes posés depuis l'époque de l'imprimerie et de la machine à écrire se sont présentés face à la technologie informatique. Les principaux problèmes sont dus à la complexité de son système graphique qui se caractérise notamment par la multitude des formes des lettres (initiale, médiane, finale et isolée), par les signes de vocalisation qui s'ajoutent au-dessus et au-dessous de chaque lettre, par les ligatures (contextuelles, linguistiques et esthétiques), par sa direction de l'écriture droite à gauche, etc.

Les deux innovations technologiques qui ont profondément marqué le domaine de l'informatique sont le standard Unicode et la norme ISO/CEI 10646. Désormais, le PBM est considéré comme le jeu de caractères universel destiné à substituer tous les jeux de caractères à 7, à 8 et à 16 bits existants. Grâce à Unicode et ISO/CEI 10646, l'industrie informatique assure la stabilité de ses données en évitant la prolifération pléthorique des jeux de caractères, augmente l'interopérabilité et l'échange de données au niveau mondial.

L'écriture arabe a été prise en compte par Unicode et ISO/CEI 10664 par 942 caractères[35]. Les caractères sont regroupés en blocs de caractères au niveau de la rangée 6, FB, FC, FD et une partie de la rangée FE. Ces blocs de caractères couvrent l'alphabet arabe de base, étendu, différentes formes des caractères, symboles religieux, ligatures de toute sorte, etc. Le standard Unicode offre une panoplie des algorithmes et des codes de formatages qui permettent un rendu fiable de l'écriture arabe avec toutes ses particularités et toutes ses fantaisies (*algorithme bidirectionnel, algorithme de mise en ordre canonique des caractères combinatoires, codes de formatage : liant et antiliant, etc.*).

5. Bibliographie

[1] ANDRE J., « Introduction : vous avez dit ligature ? », *Cahiers Gutenberg*, n° 22, septembre 1995, p. 1-4.

[2] FANTON M., « Ligatures et informatique », *Cahiers Gutenberg*, n° 22, septembre 1995, p. 61-85.

[3] HARALMBOUS Y., « Tour du monde des ligatures », *Cahiers Gutenberg*, n°22 septembre 1996, p. 87-95.

[4] FEVRIER J. G., *Histoire de l'écriture,* Paris : Editions Payot et Rivages, 1948.

[5] ISO, *ISO 9036 : information processing-Arabic 7 bits coded character set for information interchange*, Genève, ISO, 1987.

[6] ISO, *ISO 8859-6 : information processing-8 bits single-byte coded graphic character sets-part 6 : latin / arabic alphabet,* Genève : ISO, 1987.

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

[8] MEYNET R., *L'écriture arabe en question*, Beyrouth, Dar El- Machreq, 1971.

6. Sites web

ALVESTRAND H., *IETF policy an character sets and languages*, Janvier 1998, RFC : 2277. ftp://ftp.isi.edu/in-notes/rfc2277.txt

DAVIS M., *Unicode 3.1.0 : the bidirectional algorihm*, 2001-03-23. http://www.unicode.org/unicode/reports/tr9/tr9-9.html

HAPAX., *Unicode 3.1 et ISO 10646 en français*. http://iquebec.ifrance.com/hapax

BEN HENDA, M. *Vers une normalisation des pratiques de communication dans le contexte d'un multilinguisme intégral (Arabe-Latin)*. http://www.chez.com/benhenda/index_fr.htm

35. D'après la version 3.2 du standard Unicode.