SIDOP 2nd Workshop on Signal and Document Processing
March 23-24, 2012 - Hammamet - Tunisia

L'Unicode des Caractères Arabes : Etat de l'Art

Manel Daagi
Signal and Document Processing Research Group
National Engineering School of Tunis
BP 37 Belvedere 1002, Tunis, Tunisia
maneldaagi@gmail.com

Sofiene Haboubi
Signal and Document Processing Research Group
National Engineering School of Tunis
BP 37 Belvedere 1002, Tunis, Tunisia
sofiene.haboubi@istmt.rnu.tn

Résumé **—Cet article présente un aperçu de l'Unicode .Il introduit et il résume ses principales caractéristiques ainsi les problèmes engendrés par la langue arabe et les causes de l'existence de ce standard.**

Keywords;unicode;ecriture arabe;analyse contextuel;affichage bidirectionnel;ligature;arabic-forms-B; arabic-forms-A;

I. INTRODUCTION

Le standard Unicode se définit comme un système de codage mettant en œuvre un mécanisme cohérant et universel de codage des caractères .Il permet de pallier aux problèmes pour chaque langue. Il admet aux textes multilingues de coexister.

Ainsi il vient de résoudre les problèmes des caractères arabes grâce à ces algorithmes : L'algorithme d'analyse contextuelle qui permet de traiter les problèmes de ligatures, ainsi il définit la forme correcte de chaque caractère selon leur position dans un mot.

De plus, l'algorithme d'affichage bidirectionnel proposé par Unicode définit le sens d'écriture des caractères arabe.

Nous avons étudié l'existence de l'Unicode, les caractéristiques et les problèmes engendrés par la langue arabe. Et les formes d'encodage qui permet de représenter les caractères d'un jeu de caractères codés.

II. POURQUOI UNICODE

Durant plusieurs années, les ordinateurs n'utilisaient que les 26 lettres de l'alphabet latin dans sa version anglaise. Les incompatibilités entre les claviers ont émergée : des incompatibilités concernant les codes pages, des problèmes de transmission électronique des textes, etc.

Par exemple, le système de codage ASCII standard ne reconnaissait pas les accents du français. Ce standard a montré son insuffisance, puisqu'il fonctionne sur 8 bits, c'est à dire qu'il ne permet que 128 positions de codage. Actuellement, il y a plus d'un million de caractères dans le monde entier qui ont besoin de codage, pour satisfaire les demandes croissantes des langues industrielles, et pour permettre à d'autres langues comme l'arabe et ses règles de liaisons d'apparaître sur un écran informatique.

Les jeux de caractères utilisés possédaient des architectures très différentes les uns des autres [2]. Pour plusieurs, la simple détection des octets représentant un caractère était un processus contextuel complexe. Les jeux de caractères classiques ne pouvaient au mieux prendre en charge que quelques langues.

La prise en charge de plusieurs langues à la fois était difficile, voire impossible. Aucun jeu de caractères ne fournissait toutes les lettres, les signes de ponctuation et les symboles techniques en usage courant utilisés pour une seule langue.

Ces problèmes ont obligés les constructeurs d'ordinateurs de créer un autre standard de codage, qui peut supporter ce nombre énorme de caractères, mais compatible avec les normes existantes, c'est la norme Unicode ou bien UCS.

III. UNICODE

Le standard Unicode a été créé par un groupe de constructeurs d'ordinateurs en 1989[2]. Il permet de définir le codage pour la majorité des caractères utilisés par les langues du monde. Chaque caractère Unicode est associé à un point de code. Les points de code Unicode sont notés sous la forme U+nnnn, où nnnn est l'hexadécimal de point de code, ou sous forme d'une chaîne de texte descriptive.

Il définit d'une manière cohérente le codage des textes multilingues [3] et facilite l'échange de données textuelles.

Grâce à Unicode, l'industrie informatique peut assurer la pérennité des données textuelles tout en évitant l'augmentation de jeux de caractères et l'interopérabilité des données.

L'Unicode simplifie le développement de logiciels et en réduit les coûts. En effet, Unicode permet de coder tous les

Identify applicable sponsor/s here. If no sponsors, delete this text box. (sponsors)

caractères utilisés par toutes les langues écrites du monde (plus d'un million de caractères sont réservés à cet effet). Tous les caractères, quelle que soit la langue dans laquelle ils sont utilisés, sont accessibles sans aucune séquence d'échappement.

Le codage de caractère Unicode traite les caractères alphabétiques, les caractères idéographiques et les symboles de manière équivalente.

L'Unicode ajoute des règles de collation, de normalisation des formes, de bidirectionnalité et de mise au point d'algorithmes standards utilisant ces propriétés.

A. Algorithme d'affichage bidirectionnel

Il est inspiré d'une solution complète proposé par le standard Unicode destiné à traiter les complications de l'écriture de la langue arabe. Il définit pour chaque caractère le sens de son écriture pour gérer les bris de texte mixte (arabe et latin).

Ainsi il permet de remplacer les caractères arabes par leurs glyphes corrects [2].

B. Algorithme d'Analyse Contextuelle

Cet algorithme proposé par le standard Unicode. Destinée essentiellement à traiter le problème de la liaison qui se trouve entre les caractères arabes et définir leur forme correcte, quel que soit le type des caractères voisins, arabes ou bien autres.

Il permet aussi de résoudre le problème de la ligature arabe LAM-ALEF, et ceci pendant l'analyse de tous les caractères [2].

C. Caractéristiques d'Unicode

Le standard Unicode fut donc conçu pour être [4]:

- *Universel :* Le répertoire doit être suffisamment étendu pour comprendre tous les caractères susceptibles d'être utilisés dans les échanges de textes habituels, y compris les principaux jeux de caractères internationaux, nationaux ou industriels.
- *Efficace :* Le texte brut doit être facile à analyser ; les logiciels ne doivent pas maintenir une variable d'état ou rechercher des séquences d'échappement, la synchronisation de caractère à partir de n'importe quel point dans le flux de caractères doit être rapide et non ambigu.
- *Uniforme :* Un jeu de caractères de largeur fixe permet de trier, de repérer, d'afficher et d'éditer des textes efficacement.
- *Non ambigu :* Toute valeur de 16 bits représente toujours le même caractère.

IV. CARACTERISTIQUES ET PROBLEMES DE LA LANGUE ARABE

Les caractéristiques qui distinguent la langue arabe des autres langues sont :

- Avec ses 28 caractères de base la langue arabe possède 78 formes graphiques, elle s'inscrit dans son intégralité graphique des caractères dits diacritiques.
- Admet un aspect cursif c'est-à-dire les différents caractères formant le mot sont liées entre eux ceci cause souvent des problèmes au niveau de l'affichage.les lettres peuvent prendre différentes formes (isolés, initiale, médiane et finale) selon leur position dans le mot. Une analyse contextuelle est nécessaire pour déterminer la forme appropriée [5].

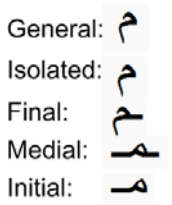

Figure 1. **Les différentes formes du caractère Min** م

- L'analyse contextuelle dans la langue arabe est encore plus complexe, sa difficulté est la présence des voyelles et autres signes diacritiques qui se place au-dessus ou au-dessous des lettres
- L'écriture et la lecture de l'arabe s'effectuent de gauche à droite. Ceci est d'ordre d'introduire des algorithmes supplémentaires pour gérer le changement du sens d'affichage ou d'impression dans les applications bilingues.
- Les formes minuscules et majuscules des caractères sont inexistantes.
- La présence des ligatures : la ligature est un glyphe spécial qui est composé de deux ou plusieurs glyphes qui sont dues à la nature cursive de l'écriture arabe [5]. Exemple : Lorsqu'un ل (forme initial ou médiane) est suivi d'un *alif* (forme finale), il faut remplacer les deux lettres par la ligature ﻻ.

V. LES TYPES D'ENCODAGE POUR LA LANGUE ARABE

Le tableau ci-dessous fournit la liste des différents types d'encodage pour la langue arabe et leur nom utilisables sur Internet et leur disponibilité sur Mac OS.

TABLE I. TYPES D'ENCODAGE POUR LA LANGUE ARABE

Type d'encodage	**Nom commun sur Internet**	**Information**
ISO 8859-6 (latin arabe)	ISO 8859-6 arabe	
Cp 864(DOS Arabic)	Cp 864	Encode les formes de présentation arabe
Cp1256 (Windows Arabe)	Windows 1256 Cp1256	Partiellement basé sur 8859-6, plus des ajouts de C1
Mac OS Arabic	X-mac6Arabic	

I. LA PLAGE DES CARACTERES ARABE DANS UNICODE

A. La présentation Forme-A (FB50-FDFF)

Elle contient des ligatures esthétiques ﷺ et linguistiques. Elle comprend aussi les codes des mots-ligatures exemple :

U+FDF0 صلے

U+FDFA ﷺ

B. La présentation forme-B (FE70-FEFF)

Elle encode les formes d'espacement, des signes diacritiques et les formes des lettres contextuelles.

Un exemple typique: les formes de positions pour les lettres arabes. Ces lettres arabes peuvent avoir jusqu'à quatre formes en fonction de leur position. Selon le concept de principe caractère et non glyphes il n'ya pas un propre code pour chaque forme qui peut prendre dans des contextes variables.

Pour des raisons historiques un nombre important des formes de présentation a été encodé en Unicode en tant que des caractères de compatibilité.

TABLE II. LES FORMES CONTEXTUELLES DU CARACTÈRE ه

Le caractère	**Les formes contextuelles des glyphes**
ه	ه ھ ھ ه
FEE9	FEEA FEEB FEEC FEE9

REFERENCES

[1] ANDRE Jacques et GOOSSENS Michel. Codage des caractères et multilinguisme : de l'ASCII à Unicode et ISO/IEC-10646. In : Cahiers GUTenberg n°20-mai 1995.pp.1-53

[2] A. ABDELHADI, L. H. Mouss and O. KADRI, "Efficient Algorithms for the integration of Arabic Language in Mobile Phone," *International Journal of Computer and Electrical Engineering, Vol. 3, No. 3, June 2011*

[3] Jacques ANDRE et Michel GOOSSENS, « Codage des caractères et multi-linguisme : de l'ASCII à UNICODE et ISO/IEC-10646, » Cahiers GUTenberg no20 — mai 1995

[4] A. ABDELHADI et O. KADRI , "L'impact Informatique de l'Intégration de la Langue Arabe dans les Téléphones Mobiles," IEEE SETIT 2009 International Conference: Sciences of Electronic, Technologies of Information and Telecommunications March 22-26, 2009 – TUNISIA

[5] M.Eddahibi, "Etude et réalisation d'outils de codageet de composition du e-document mathématique arabe," thèse octobre 2007.