ALTRALANG Journal
e-ISSN: 2710-8619 p-ISSN: 2710-7922
Volume 6 Issue 2 / December 2024 pp. 199-212
DOI: https://doi.org/10.52919/altralang.v6i2.479

Le système d'écriture de l'arabe dans le codage informatique : besoins et contraintes

MAMMERI Mahmoud Fawzi*
Ecole Supérieure de Commerce ESC Alger, Algérie
mf.mammeri@esc-alger.dz

Reçu: 21/03/2024, **Accepté:** 22/12/2024, **Publié:** 31/12/2024

Arabic Writing System in Computer Coding: Needs and Constraints

ABSTRACT: *The writing system of any language is encoded on computer in two main parts: one of which concerns alphabetical graphemes, the other, non-alphabetical graphemes (ponctuations and other typographical signs). The first part is always language specific (script). The second, given the universal character of non-alphabetical signs, is shared by all the scripts and languages of the world. This article discusses the needs of the Arabic writing system in relation to the coding of these two types of graphemes in digital systems and some of the constraints that guided its implementation in the Unicode standard. This work leads to a discussion of gaps in some aspects of this implementation and possible improvements.*

KEYWORDS: Arabic writing system, Unicode, alphabetical graphemes, non-alphabetical graphemes, grapholinguistics

RÉSUMÉ : *Le système d'écriture de toute langue est codé sur ordinateur en deux grandes parties : une partie qui concerne les graphèmes alphabétiques et une autre qui concerne les graphèmes ponctuo-typographiques. La première partie est toujours spécifique à une langue (script). La seconde, étant donnée le caractère universel des signes ponctuo-typographiques, est partagée par l'ensemble des scripts et langues du monde. Cet article parle des besoins du système d'écriture de l'arabe par rapport au codage de ces deux types de graphèmes dans les systèmes numériques et de quelques unes des contraintes qui ont conditionnées son implémentation dans le standard Unicode. Ce travail nous mène à une discussion sur des lacunes concernant certains aspects de cette implémentation et les améliorations possibles.*

MOTS-CLÉS : Système d'écriture de l'arabe, Unicode, graphèmes alphabétiques, graphèmes ponctuo-typographiques, grapholinguistique

* Auteur correspondant : **MAMMERI Mahmoud Fawzi**, mf.mammeri@esc-alger.dz

ALTRALANG Journal
Volume 6 Issue 2 / December 2024

Depuis la naissance de l'imprimerie moderne en Europe, la prise en charge du caractère arabe a toujours eu un important retard par rapport au caractère latin. Ceci est dû au fait que les outils et technologies autour du caractère ont toujours été développés pour le compte du latin : d'abord avec la typographie (composition mobile utilisant le métal), ensuite la dactylographie (écriture mécanique), la typographie moderne (photocomposition ou composition photographique) et enfin le numérique (codage informatique des caractères) ; à ceci s'ajoute le fait que toutes ces technologies trouve leur origine dans les caractères utilisés par les graveurs pour les inscriptions lapidaires ensuite l'écriture manuscrite. Ce retard qui se comptait en siècles avec l'imprimerie, à cause de l'hégémonie de la religion[1], a de l'être réduit en dizaines d'années avec le numérique, le pouvoir ayant eu basculé du côté de la socio-économie[2].

Il aura fallu plusieurs dizaines d'années de recherche et de développement, principalement dans le secteur de l'industrie mécanique ensuite du numérique, pour permettre une prise en charge acceptable du caractère arabe dans le domaine du numérique : tout d'abord, avec le développement des dispositions pour claviers et des mécaniques nécessaires au fonctionnement de la machine à écrire arabe, ensuite avec le développement des codages informatiques modernes et surtout des outils logiciels pour un meilleur rendu du texte à travers de nouveaux algorithmes de connexion des caractères arabes.

Le texte numérique a rencontré, à ses débuts, plusieurs problèmes dont le plus important est le codage des caractères et la prise en charge des scripts et langues du monde. La plupart de ces problèmes ont été réglés avec la publication du standard Unicode en 1991.[3] C'est ce qui a permis un essor considérable dans le domaine du texte et de la typographie modernes.[4] En effet, le progrès de l'informatique dans ses deux aspects matériel et logiciel a donné naissance à de nouvelles possibilités graphiques notamment des outils spécialement dédiés au design typographique, ou création de polices de caractères, encore inconnus il y a quelques décennies. Ce qui était difficile à réaliser et parfois à imaginer pendant quatre siècles d'imprimerie est aujourd'hui possible et à moindre coût avec l'ordinateur, les nouvelles technologies de l'information et de la communication et le codage informatique universel. Les développements se sont multipliés à tel point que l'on peut mettre des textes de différentes langues et de différents scripts dans un même document numérique.

Aujourd'hui, les textes, quels que soient leurs genres, sont presque tous plurilingues[5], ceci est principalement du au phénomène de citation qui consiste à citer des passages de textes dans des langues autres que la langue principale du texte. Ceci prend la forme de citations à l'intérieur du texte et de références bibliographiques à la fin. Du point de vue du programmeur, ces nouveaux textes ont donné lieu à de nouvelles problématiques dans le traitement automatique des chaines de caractères et des textes. Du point de vue du scripteur, si le numérique a permis de trouver des solutions aux difficultés de la machine à écrire et de l'imprimerie, avec une offre de plus de 140.000 caractères (toutes langues et scripts confondus), il a ramené avec lui son lot de difficultés quant à la maitrise et l'accès à cette large collection de caractères. Ceci nécessite en quelque sorte une nouvelle forme de littéracie dans le domaine de la saisie et la manipulation des textes dans le numérique : la familiarisation avec cette nouvelle offre de caractères et la maitrise du clavier et des outils logiciels qui lui sont complémentaires.

Un texte plurilingue dont la langue principale est l'arabe peut contenir, comme tous types de textes plurilingues, d'une part, des lettres de l'alphabet arabe, des ponctuations, des diacritiques, des symboles

[1] L'église s'est emparée des technologies du texte depuis l'invention de l'imprimerie et ceci pendant quatre siècles (1450-1850).
[2] Les politiques s'intéressent de plus en plus à ouvrir de nouveaux marchés pour imposer leurs modèles socio-économiques.
[3] Unicode est un standard du Consortium Unicode conforme à la norme ISO-10646 de l'ISO.
[4] Le caractère n'est plus uniquement utilisé dans l'écriture des textes mais aussi dans la marque pour accentuer l'image de marque des entreprises : des marques très prestigieuses s'identifient facilement grâce à des logos minimalistes et une fonte qui leur est propre.
[5] Avec Unicode, un texte dans une langue, écrit avec un script donné, est plurilingue du moment où il emprunte des caractères d'un autre script.

(mathématiques, monétaires, etc.), des signes typographiques (espaces, tirets, fin de paragraphe, etc.), des diacritiques et lettres utilisés en translitération, et d'autres part, des citations ou des textes parallèles dans d'autres langues, qui nécessitent à leur tour des lettres, des diacritiques, des ponctuations, etc.

Pour réaliser une quelconque production en arabe, il est important de veiller à composer des documents de bonne qualité typographique. Ceci n'est pas toujours facile à réaliser et le scripteur est souvent confronté à des difficultés en manipulant des textes plurilingues. Pour donner une idée sur les problèmes que peut rencontrer le scripteur dans la réalisation de tels textes, nous pouvons citer la difficulté de maitriser le sens d'écriture en un point donné du texte, l'absence de certains caractères importants sur la quasi-totalité des claviers des ordinateurs, la similarité entre certains caractères tels que le trait d'union avec le tiret ou la non maitrise des espaces et des traits d'union en fin de lignes. Tous ces problèmes rendent difficile le choix du scripteur entre tel ou tel signe.

La question principale à laquelle nous tenterons d'apporter des éléments de réponse pourrait être reformulée de la manière suivante : le scripteur, a-t-il tous les outils nécessaires, en termes de caractères et d'accès à ces caractères, pour écrire un texte numérique de bonne qualité typographique ? La réponse à cette question nous amène à discuter un certains nombres de questions secondaires : qu'est ce qu'un texte numérique ? qu'est ce qu'un texte numérique de bonne qualité typographique ? quels sont les outils offerts par le système d'écriture de l'arabe pour écrire du texte numérique arabe ? comment a été codé l'arabe dans le codage universel ? le codage de l'arabe a-t-il respecté le principe d'universalité de certains signes du système d'écriture de l'arabe, à savoir les graphèmes ponctuo-typographiques ?.

La réponse à cette problématique contribue à lever beaucoup de questionnements sur la typographie de l'arabe standard moderne et à doter le scripteur de connaissances en typographie de l'arabe qui ne fait pas encore l'objet de normes communes. Nous discutons aussi quelques problèmes liés aux traitements automatiques de certains caractères du script arabe.

1/ Matériels et méthodes

Ce travail s'intéresse à l'écriture des textes arabes – ces textes étant par nature numériques plurilingues et codés en Unicode – et leur conformité aux normes typographiques modernes. Pour la réalisation de ce travail, nous avons parcouru la documentation en rapport avec le thème de notre étude. Notre thème d'étude concerne deux domaines essentiels : le texte numérique basé sur le codage informatique Unicode et le système d'écriture de l'arabe.

Texte numérique. Tout texte composé sur un ordinateur – qu'il soit sur une page Web, dans un logiciel de traitement de texte, etc. – n'est autre qu'un ensemble de valeurs numériques. Une suite de caractères telle que le mot « texte » est codée par les nombres 116, 101, 120, 116 et 101 qui représentent respectivement les caractères bas de casse « t », « e », « x », « t » et « e ». C'est ces valeurs numériques qui sont encodées puis stockées dans les mémoires des ordinateurs à la place des caractères alphabétiques ; et c'est la raison pour laquelle qu'on parle de texte numérique. Le premier codage informatique qui a été utilisé pour le codage des textes est le code ASCII apparu en 1969. Depuis, plusieurs codes l'ont suivi pour améliorer la prise en charge des alphabets des différentes langues du monde. Ces différentes améliorations successives ont conduit à l'émergence d'un codage universel qui satisfait toutes les langues et scripts du monde : c'est le codage universel plus connu sous le nom Unicode. Après la publication de la première version du standard Unicode en 1991, les systèmes pour l'édition des textes numériques se sont tous convertis dans le codage universel. Depuis, tous les nouveaux textes numériques sont codés en Unicode. La source la plus complète sur le standard Unicode est incontestablement le site du Consortium Unicode[6] ; en plus d'être complète, cette ressource est mise à jour à chaque nouvelle version du standard.

[6] https://home.unicode.org/

Système d'écriture de l'arabe. Le système d'écriture de l'arabe est un Abjad ou alphabet consonantique, dans lequel les sons consonantaux sont représentés comme des caractères de base. Quant aux voyelles, elles sont représentées de deux manières différentes : les voyelles longues par des graphèmes (caractères de base) tandis que les voyelles courtes sont notées en diacritiques sur leurs supports consonantiques. Ce système de base permet d'écrire les mots de la langue, mais il est insuffisant pour rassembler ces mots en phrases, paragraphes et textes. Pour écrire des textes en arabe, deux sortes de signes – ou graphèmes sont nécessaires : des graphèmes alphabétiques, utilisés pour représenter les phonèmes, et des graphèmes ponctuo-typographiques, qui ne correspondent à aucun phonème mais qui servent à organiser le texte ou à se lui substituer. La particularité du système d'écriture de l'arabe est qu'il est cursif, c.-à-d. que chaque graphème alphabétique peut avoir plusieurs formes selon le contexte où il apparait : initiale, médiale, finale ou isolée[7]. Même si ces variations contextuelles (dessins des glyphes) ne sont pas codées en Unicode, elles sont prises en charge par les moteurs de rendus. Une complexité additionnelle du script arabe, dans sa forme numérique, est sa bi-directionalité. Tandis qu'il est traditionnellement qualifié de script droite-à-gauche, dans sa version numérique le script arabe est bi-directionnel. Ceci est du au fait que certains caractères Unicode, en particulier les chiffres, font partie intégrante du script latin et sont orientés de gauche-à-droite quel que soit le script avec lequel ils sont utilisés.

La collecte des données depuis les sources consultées a donné lieu aux résultats exposés dans la section subséquente.

2/ Résultats : Système d'écriture arabe et codage informatique

2.1/ Alphabet, formes et ligatures

L'écriture arabe possède plusieurs caractéristiques qui rendent son traitement numérique plus complexe, ce qui nécessite plusieurs outils supplémentaires pour la composition des textes. Il s'agit d'une écriture :

À script. Elle possède un certain nombre de lettres. On en compte 29 lettres : ء, ا, ب, ت, ث, ج, ح, خ, د, ذ, ر, ز, س, ش, ص, ض, ط, ظ, ع, غ, ف, ق, ك, ل, م, ن, ه, و, ي. Comme il est très facile de le remarquer sur les graphies de ces lettres, le modèle de l'écriture arabe utilise 18 graphèmes de base (sans compter la graphie de la *hamza*) et la notion de *'i'jām*[8] — qui consiste à rajouter un, deux ou trois points sus- ou souscrits sur un graphème donné — pour noter les 29 lettres (consonnes et voyelles longues) utilisée dans la langue. Ces lettres sont implémentées dans le bloc Unicode x0600 (respectivement : U+0627-U+0628, U+062A-U+063A, U+0641-U+0648, U+064A). Pour assurer une composition numérique des lettres arabes, les systèmes d'édition ont besoin de polices de caractères. Plusieurs polices Unicode existent actuellement pour l'arabe : des polices monolingues (p. ex. Arabic Typesetting, Simplified Arabic ou Traditional Arabic), qui sont les plus utilisées — actuellement — dans les textes arabes écrits sous Windows, mais également d'autres qui sont soit multilingues (p. ex. Times New Roman), soit universelles (p. ex. Arial Unicode MS). Certaines polices sont plus robustes que d'autres, donnant plus de possibilités pour le rédacteur et ainsi facilitant le traitement des textes par la suite.

Basée sur un abjad[9]. C'est une écriture qui utilise un alphabet consonantique représenté en Unicode par un certain nombre de caractères abstraits dits caractères de base. Pour représenter ses voyelles courtes,

[7] Lorsqu'il est isolé, un graphème alphabétique n'est en contact avec aucun autre graphème. Dans les trois autres cas, il est connecté soit au suivant soit au précédent soit aux deux graphèmes qui lui sont adjacents.

[8] C'est cette même particularité de l'écriture arabe (*'i'jām*) qui a permis d'étendre l'alphabet arabe à d'autres langues telles que le persan, l'ourdou, le turc ottoman, et de manière historique, les langues turques d'Asie centrale.

[9] *Abjad* est la transcription courante du mot arabe أَبَجَد. Ce dernier a été composé à partir des quatre premières lettres de l'alphabet arabe selon l'ordre primaire commun à tous les alphabets sémitiques qui ont comme ancêtre le phénicien : phénicien (*Aleph*, *Beth*, *Ghimel*, *Daleth*), arabe (*Alef*, *Beh*, *Jeem*, *Dal*), hébreu (*Alef*, *Bet*, *Gimel*, *Dalet*), Syriaque (*Alaph*, *Beth*, *Gamal*, *Dalath*), etc. Un *abjad* (ou alphabet consonantique) est un alphabet qui ne comporte que des consonnes, ce qui veut dire que les voyelles ne sont pas distinguées au niveau de l'alphabet. Les voyelles — qui demeurent prononcées à l'oral — sont le plus souvent

elle a donc besoin d'un ensemble de caractères supplémentaires sans chasse dits caractères combinatoires, destinés à s'afficher collés aux caractères de base. Le système d'écriture arabe utilise huit caractères combinatoires notés :[10] ◌ً, ◌ٌ, ◌ٍ, ◌َ, ◌ُ, ◌ِ, ◌ّ, ◌ْ. Ces caractères sont tous localisés dans le bloc Unicode x0600 et sont codés respectivement avec les points de code 064B-0652. Pour afficher le glyphe « وُ », par exemple, nous aurons besoin d'insérer le caractère de base « و » (U+0648 ARABIC LETTER WAW) suivi du caractère combinatoire « ◌ُ » (U+064F ARABIC DAMMA). Les caractères combinatoires peuvent aussi être empilés : un caractère de base déjà combiné avec un caractère combinatoire « وُ » peut être à son tour à nouveau combiné à un autre caractère combinatoire « ◌ّ » (U+0651 ARABIC SHADDA) pour donner lieu à la graphie « وُّ ».

Cursive. L'écriture arabe, qu'elle soit manuelle ou sur ordinateur, n'est pas réalisée par la simple juxtaposition de lettres — comme c'est le cas général pour l'écriture latine sur ordinateur si l'on n'en prend pas en compte les cas de ligature — mais nécessite un nombre plus important de graphies. En effet, le fait qu'elle soit cursive veut dire que les lettres arabes prennent des formes selon le contexte où elles apparaissent dans le flux d'affichage. Ces formes peuvent être classifiées en deux classes : des ligatures contextuelles (ou formes contextuelles) et une ligature linguistique unique *lām-alif*. Une lettre telle que ت peut s'écrire d'une manière liée soit au début (تـ) soit au milieu (ـتـ) soit à la fin (ـت) d'un mot, tout comme elle peut apparaître isolée (ت). Selon les formes qu'elles peuvent prendre dans le contexte, les lettres arabes se divisent en six groupes :

- 22 des 29 lettres présentent les quatre formes possibles – initiale, médiane, finale et isolée – : ب, ت, ث, ج, ح, خ, س, ش, ص, ض, ط, ظ, ع, غ, ف, ق, ك, ل, م, ن, ه, ي ;
- six lettres présentent seulement les deux formes finale et isolée : ا, د, ذ, ر, ز, و ;
- la *hamza* qui présente des formes très particulières : elle peut prendre une forme isolée (ء), participer à la formation de caractères diacritiques (آ, أ, ؤ, إ, ئ, ـآ, ـأ, ـؤ, ـإ, ـئ, ـئـ) ou faire partie de certaines des réalisations de la ligature *lām-alif* (ﻵ, ﻷ, ﻹ, ﻻ, ﻸ, ﻼ) ;
- le *alif* qui présente deux formes auxiliaires, l'une, utilisée uniquement dans les textes coraniques : c'est la voyelle longue *alif qaṣīrat* (◌ٰ), l'autre, pour représenter une prononciation particulière du *alif* : c'est la voyelle longue *alif maqṣūra* (ى) ;
- la *tā' marbūṭat* qui présente une occurrence spéciale sous une forme fermée : *tā' marbūṭat* (ة) ;
- la ligature *lām-alif* (ﻻ, ﻼ).

De toutes ces formes, uniquement les six formes de la hamza non concernées par la ligature *lām-alif* (U+0621-U+0626), les deux formes du *alif* (*alif maqṣūra* : U+0649 et *alif qaṣīrat* : U+0670) et la forme isolée de la *tā' marbūṭat* (U+0629) sont implémentées dans le bloc Unicode x0600. Les autres, à savoir les différentes ligatures pour les 22 lettres à quatre formes contextuelles, les 6 lettres à deux formes contextuelles et la *tā' marbūṭat* et les ligatures *lām-alif* (qu'elles soient porteuses ou non de *hamza*), sont implémentées dans le bloc Unicode *Arabic Presentation Forms-B*, mais leur utilisation est déconseillée par Unicode ; elles ont été rajoutées à Unicode pour des besoins de compatibilités avec certains codages antérieurs. Ainsi, ces formes sont prises en charge directement au niveau du moteur de rendu, et le rédacteur n'a pas à s'en soucier : il lui suffit d'insérer des lettres et c'est au moteur de rendu de calculer les liaisons nécessaires[11] et, par conséquent, d'utiliser les glyphes appropriés à partir des tables de la police de caractères utilisée.

ignorées à l'écrit. Si elles ne le sont pas, elles sont optionnellement indiquées par le biais de marques secondaires — sous la forme de diacritiques — sur ou sous les consonnes ; c'est le cas, en arabe, des manuels scolaires et du coran qui sont totalement ou partiellement diacritisés.

[10] Pour les différencier des autres caractères : les caractères combinatoires sont représentés sur un cercle en pointillés.

[11] Pour réaliser ce calcul, les moteurs de rendu font appel à un fichier spécial *ArabicShaping.txt* disponible sur le site d'Unicode à l'adresse https://www.unicode.org/Public/13.0.0/ucd/ArabicShaping.txt.

Orientée de droite à gauche. Par défaut, le sens d'écriture du latin est pris comme référence dans la gestion d'affichage des scripts – ce sens n'a rien à voir avec le sens logique des caractères en mémoire. Par conséquent, un moteur de rendu a toujours besoin d'algorithmes supplémentaires pour gérer la directionalité des textes qui s'écrivent de droite à gauche, notamment les textes arabes. Il s'agit d'un ensemble d'algorithmes permettant de gérer tous les scripts de droite à gauche connu sous le nom d'Algorithme BiDi.

Monocamérale. Elle n'utilise pas de casse capitales/minuscules. En d'autres termes, les débuts des phrases, les noms propres et les abréviations ne sont pas marqués, ce qui rend leurs traitements automatiques plus complexes.

2.2/ Outils pour la ponctuation, la typographie et les nombres

Pour composer des textes arabes, le système d'écriture arabe a besoin d'un certain nombre d'outils supplémentaires qui ne sont pas nécessairement spécifiques à l'arabe : ponctuations, chiffres, signes typographiques et symboles (mathématiques, monétaires, linguistiques, musicaux, etc.).

Ponctuation arabe dans Unicode

L'arabe, comme toutes les langues naturelles, a reconduit la ponctuation universelle, mais avec une légère adaptation liée essentiellement à l'orientation de son écriture qui est à l'inverse du latin de droite à gauche. Les signes de ponctuation proposés pour le texte arabe dans le cadre du standard Unicode ne sont pas tous spécifiques au script arabe et la plupart d'entre eux sont unifiés avec la ponctuation latine : on parle ainsi de ponctuation générale. La ponctuation générale est codée en Unicode avec un principe d'économie (c.-à-d., de non duplication) et est localisée dans le Plan Multilingue de Base (0000 : FFFF). Elle est commune à toutes les langues et occupe des positions dans les blocs *Basic Latin*, *Latin-1 Supplement* et *General Punctuation*. Nous pouvons catégoriser les ponctuations utilisées en arabe en deux types : (i) des signes issus de la ponctuation générale avec un même glyphe et (ii) des signes issus de la ponctuation générale avec un glyphe différent.

Signes issus de la ponctuation générale avec un même glyphe

Les textes numériques arabes utilisent les ponctuations standards :

- le point d'exclamation (EXCLAMATION MARK : 0021)
- le point (FULL STOP : 002E)
- la barre oblique ou transversale (SOLIDUS : 002F)
- les deux points (COLON : 003A)
- le tiret : trait d'union (HYPHEN : 2010), tiret court (FIGURE DASH : 2012), tiret moyen (EN DASH : 2013) et tiret long (EM DASH : 2014)
- les points de suspension (HORIZONTAL ELLIPSIS : 2026)[12]
- les parenthèses : parenthèse gauche (LEFT PARENTHESIS : 0028) et parenthèse droite (RIGHT PARENTHESIS : 0029)
- les crochets : crochet gauche (LEFT SQUARE BRACKET : 005B) et crochet droit(RIGHT SQUARE BRACKET : 005D)
- les accolades : accolade gauche (LEFT CURLY BRACKET : 007B) et accolade droite (RIGHT CURLY BRACKET : 007D)
- les guillemets : quottes (QUOTATION MARK : 0022), guillemet angle double vers la gauche ou guillemet français gauche (LEFT-POINTING DOUBLE ANGLE QUOTATION MARK : 00AB), guillemet angle double vers la droite ou guillemet français droit (RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK : 00BB), guillemet simple gauche (LEFT SINGLE QUOTATION MARK : 2018), guillemet simple droit (RIGHT SINGLE QUOTATION MARK : 2019), guillemet

[12] Ce n'est que depuis peu de temps que ce caractère est pris en charge par les logiciels de traitements des textes arabes, à la place des trois points successifs sous la forme des trois glyphes U+002E.

simple bas ressemblant au chiffre 9 (SINGLE LOW-9 QUOTATION MARK : 201A), guillemet simple haut ressemblant au chiffre 9 inversé (SINGLE HIGH-REVERSED-9 QUOTATION MARK : 201B), guillemet double gauche (LEFT DOUBLE QUOTATION MARK : 201C), guillemet double droit (RIGHT DOUBLE QUOTATION MARK : 201D), guillemet double bas ressemblant au chiffre 9 (DOUBLE LOW-9 QUOTATION MARK : 201E), guillemet double haut ressemblant au chiffre 9 inversé (DOUBLE HIGH-REVERSED-9 QUOTATION MARK : 201F), guillemet angle simple vers la gauche (SINGLE LEFT-POINTING ANGLE QUOTATION MARK : 2039) et guillemet angle simple vers la droite (SINGLE RIGHT-POINTING ANGLE QUOTATION MARK : 203A)

Certains des signes sus-énumérés sont utilisés avec une spécificité de la typographie arabe : les guillemets, crochets et accolades droits sont utilisés comme étant des signes ouvrants et les guillemets, crochets et accolades gauches comme étant des signes fermants. C'est ainsi qu'une parenthèse ouvrante en écriture GD (latine par exemple) est fermante en écriture DG et une parenthèse fermante en écriture GD est ouvrante en écriture DG. Ainsi, pour l'orientation des ponctuations commune, le calcul de la direction en Unicode se fait selon la règle suivante : une ponctuation commune est dépourvue de toute directionalité implicite et c'est l'algorithme bidirectionnel d'Unicode (BiDi) qui permet de calculer sa direction à partir de la direction du texte rendu. Par conséquent, en pratique, un guillemet ouvrant, par exemple, est obtenu sous MS Word par la combinaison des touches Alt+0171 qu'il soit utilisé dans un texte latin (Gauche à Droite) ou dans un texte arabe (Droite à Gauche).

Signes issus de la ponctuation générale avec un glyphe différent

Le texte arabe utilise aussi d'autres signes de ponctuation universels avec une graphie légèrement modifiée par rapport à celle utilisée pour le latin mais avec la même sémantique. Le recours à une telle apparence est surtout guidé par des besoins d'esthétique et de lisibilité du texte. Ces signes sont codés indépendamment dans le bloc arabe de base :

- la virgule « ، » (ARABIC COMMA : U+060C)
- le point virgule « ؛ » (ARABIC SEMICOLON : U+061B)
- le point d'interrogation « ؟ » (ARABIC QUESTION MARK : U+061F)

Ces caractères de ponctuation spécifiques ont une directionalité fixe (Droite à Gauche). (The Unicode Consortium, 2011)

Autres signes typographiques

Outre l'espace justifiante utilisée pour la justification et la mise en forme du texte numérique, le texte arabe utilise d'autres ponctuations pour (ou d'aide à) la mise en forme du texte. Dans ce cadre, Unicode propose les caractères suivants souvent utilisés :

- la *kashida* (ARABIC TATWEEL : U+0640)
- l'espace insécable (NO-BREAK SPACE : 00A0)
- le point médian (MIDDLE DOT : 00B7)
- le signe degré (DEGREE SIGN : 00B0)
- le signe du paragraphe (PILCROW SIGN : 00B6)
- le signe inversé du paragraphe (REVERSED PILCROW SIGN : 204B)
- Le séparateur de décimaux (ARABIC DECIMAL SEPARATOR : U+066B)
- Le séparateur de milliers (ARABIC THOUSANDS SEPARATOR : U+066C)

Tous ces caractères ont une utilisation pratique. La *kashida* est utilisée dans le texte arabe pour (i) justifier le texte en rallongeant ses caractères – on l'oppose à la justification utilisant l'espace (ou blanc typographique) dans le script latin ; l'espace pouvant être souvent plus ou moins élargie entre des mots ou, moins souvent, entre des caractères d'un même mot – ou (ii) porter des diacritiques (ou *taṣkīl*) en l'absence d'un caractère de base.

En pratique, dans les logiciels d'édition tels que MS Word, le point médian « · » est inséré dans un document à la place des espaces justifiantes pour vérifier (i) qu'il n'y a pas d'espaces en surplus et pour (ii)

les différencier des espaces insécables, qui, elles, sont représentées par le signe *degré* « ° ». De même, les deux signes du paragraphe servent à comptabiliser et gérer les paragraphes. Tous ces caractères sont des caractères non imprimables et n'apparaissent que pour des opérations de mise en forme. Dans le logiciel MS Word, on peut les afficher et contrôler son texte à travers l'onglet « Affichage » du menu « Options Word » (voir Fig. 1).

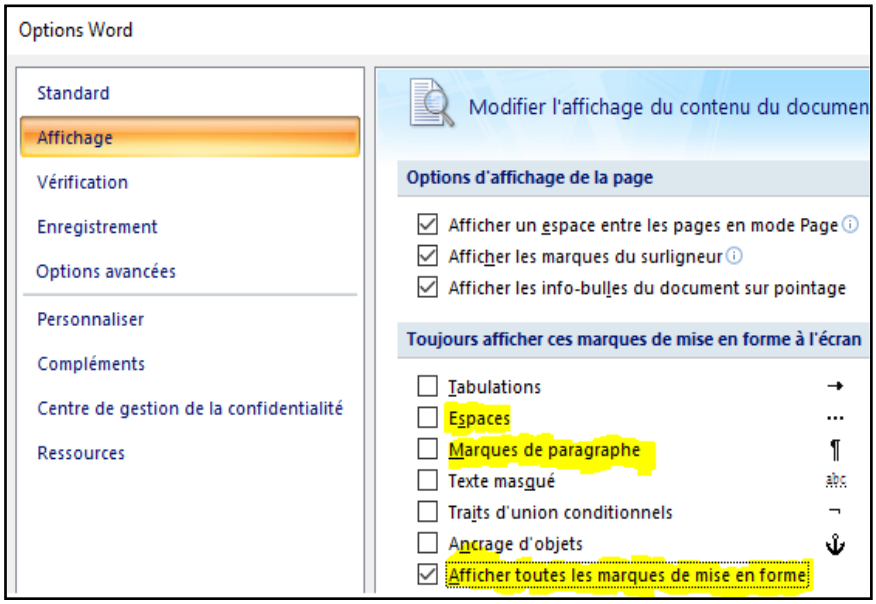


Figure 1. Affichage des marques de mise en forme dans MS Word

Enfin, les séparateurs de décimaux et de milliers sont utilisés avec les nombres et servent à les aérer pour plus de lisibilité.

Chiffres

Il existe deux types de chiffres utilisés dans les textes arabes :

- les chiffres arabes sous leur forme hindi : ٠, ١, ٢, etc., utilisés en Orient ;
- les chiffres arabes sous leur forme occidentale : 0, 1, 2, etc., utilisés dans les pays du Maghreb.

Les chiffres hindis sont situés dans le bloc « Arabic » (0660-0669), tandis que les chiffres occidentaux sont situés dans le bloc « Basic Latin » (0030-0039). Malgré le faible typage de ces caractères, leur orientation est toujours de gauche à droite et ce quel que soit le script utilisé.

Symboles (mathématiques, monétaires, etc.)

En plus des caractères alphabétiques, les ponctuations, les chiffres et certains signes typographiques, le texte arabe peut avoir besoin d'un certain nombre de symboles supplémentaires. Ces symboles, qui portent une sémantique et font partie de l'information véhiculée par le texte – à l'inverse de la ponctuation[13] –, sont généralement utilisés comme :

[13] Le point, par exemple, indique la fin d'une phrase : s'il est omis, certes le texte est difficile à lire mais reste toujours lisible ; le texte véhicule toute son information. Par contre, l'omission d'un caractère tel que « $ » (Dollar) peut engendrer une ambigüité lors de la lecture d'un texte contenant des sommes d'argent et des monnaies.

- des abréviations : le signe « $ » pour remplacer l'expression « US »[14], « ٪ » (Pourcentage), « ؉ » (Pourmille), « € » (Euro), « © » (Copyright), « ℗ » (Copyright phonographique), « ° » (Degré), « °C » (Celsius), etc. ou
- des fonctions particulières : « # » (Croisillon :[15] pouvant jouer le rôle d'un Hashtag dans les réseaux sociaux, indiquer un numéro dans une adresse en typographie américaine, etc.), « + » (signe « plus » : addition de nombres, concaténation de chaînes de caractères dans les languages de programmation, etc.), « − » (signe « moins » : soustraction, marque du signe d'un nombre négatif et l'opposé d'un nombre), « @ » (Arobase : surtout utilisé en informatique dans la messagerie, les réseaux sociaux et les langages de programmation), « * » (Astérisque : très utilisé en typographie, linguistique, mathématiques, informatique, etc. pour faire un renvois vers une note de bas de page, en énumération à la place des tirets et des puces, etc.), « & » (Esperluette : joindre plusieurs auteurs dans une citation, etc.), un chiffre suscrit ou souscrit (exposant ou indice : pour référencer un texte comme étant une référence en bas ou en fin de page, etc.), etc.

Les symboles utilisés dans le texte arabe peuvent appartenir à différents blocs : Latin, Arabic, Arabic Mathematical Alphabetic Symbols[16] (U+1EE00–U+1EEFF), etc. Le seul souci réside dans la direction du caractère utilisé qui est déterminée par la propriété Unicode *Bidi Class*.

Dans la section subséquente, nous présentons une discussion sur certains aspects de l'implémentation Unicode des caractères utilisés dans le texte arabe. Cette discussion porte aussi sur les difficultés rencontrées par les scripteurs dans la manipulation de certains caractères.

3/ Discussion : Difficultés et problématiques du texte arabe

D'après les résultats que nous venons d'exposer, les graphèmes alphabétiques et ponctuo-typographiques utilisés dans le texte arabe sont tous pris en charge dans le codage universel soit comme étant des signes spécifiques à l'arabe, codés dans les blocs réservés à l'arabe, soit comme des signes universels, codés dans les blocs réservés au script latin. L'accès à tous ces signes n'est pas possible uniquement avec le clavier. Le scripteur n'étant pas toujours en mesure d'utiliser le bon signe, recoure souvent à des variantes ou même des signes alternatifs. La solution à ce problème est résolue par l'utilisation des outils offerts par les éditeurs et traitements des textes, comme l'outil Symbole et les outils de correction automatique de MS Word, ou par les systèmes d'exploitation, comme la Table des caractères de Windows. Faudrait-il encore savoir ce que l'on cherche comme signe pour pouvoir le localiser dans l'outil utilisé.

Une question qui nous interpelle dans la lecture et l'analyse des résultats concerne certaines ponctuations qualifiées, dans le standard Unicode, de ponctuations arabes. Ce sont la virgule (،), le point virgule (؛) et le point d'interrogation (؟) dites ponctuations arabes (ARABIC COMMA, ARABIC SEMICOLON, ARABIC QUESTION MARK).[17] Cette qualification est-elle justifiée ? Pour répondre à cette question, nous nous basons sur l'hypothèse suivante : la ponctuation (les ponctuations) est (sont) universelle (s) ; en d'autres termes, un signe de ponctuation est censé avoir la même sémantique dans toutes les langues – l'objectif étant de faciliter la communication entre langues. Ceci apparait clairement dans la

[14] Pour « United States » ; la première lettre de chacun des deux mots de l'expression – « U » et « S » – ont été superposées l'un sur l'autre et ont donné lieu au symbole : « S » avec deux barres verticales. Le signe « $ » est très utilisé en informatique ; surtout dans les langages de programmation, les expressions rationnelles, les tableurs graphiques…

[15] Le croisillon « # » (U+0023) est différent du dièse « ♯ » (U+266F), avec lequel il est souvent confondu.

[16] C'est un bloc spécifique à l'arabe réservé aux symboles mathématiques.

[17] Nous rappelons au lecteur que dans Unicode la ponctuation universelle est implémentée dans le bloc Ponctuation Générale. Cette ponctuation est utilisée par toutes les langues et scripts du monde. C'est le cas de l'arabe, sauf pour trois ponctuations à savoir la virgule, le point virgule et le point d'interrogation pour lesquelles ont été créés trois nouveaux codes dans le bloc Arabe.

célèbre lettre d'Aḥmad Zakī dans son argumentation pour l'institution des ponctuations européennes dans le texte arabe :

> وإنما جنحت إلى هذا التوفيق بين القواعد العربية وبين العلامات الأجنبية، لتوحيد العمل، وتقليل الكلفة، وتسهيل السبيل: خصوصًا أن هذه العلامات قد شاع استعمالها في المدارس والمطبوعات والمخطوطات العربية في عصرنا هذا.
>
> [...]
>
> وأهم الدواعي التي قضت بالتعويل على هذه العلامات، أن التلاميذ المصريين في جميع المدارس الأميرية والأهلية والأجنبية يتعلمون هذه العلامات، أثناء تلقيهم اللغات الأجنبية. فلو اخترت علامات أخرى، لكان ذلك العمل موجبًا للتهويش (التشويش) على الطلبة، ولا سيما حديثي العهد منهم بالدراسة. وفي ذلك ما فيه، مما يتحتم تلافيه.
>
> (Aḥmad Zakī, 2013 : 11)[18]

À ceci près que les signes choisis pour ces ponctuations doivent s'assortir esthétiquement avec le caractère arabe qui est orienté de droite à gauche. Aḥmad Zakī parle aussi de la flexibilité du mouvement de la main lors de l'écriture du signe :

> فلهذه الأسباب كلها، رأيت وجوب الاعتماد على هذه العلامات، بعد تعديل وضعها، بحيث يمكن كتابتها بالقلم العربي: مراعاةً لحركة اليد في الكتابة، من اليمين إلى اليسار.
>
> (Aḥmad Zakī, 2013 : 11)[19]

Cependant, le fait que le signe de ponctuation doit s'assortir avec le texte arabe ne justifie pas le recourt à un nouveau signe de ponctuation spécifique.

L'universalité des signes de ponctuation a été matérialisée dans le numérique dans tous les codages informatiques postérieurs au code ASCII. Dans une version modifiée de la norme ISO 646[20], adaptée à la langue arabe, les signes de ponctuation utilisés dans le texte arabe ont gardé les mêmes positions (donc les mêmes codes) que ceux du latin[21] ; seul le dessin du signe (glyphe) est modifié si nécessaire – c'est le cas des trois signes virgule, point virgule et point d'interrogation. (Voir Fig. 2.)

---

[18] « J'ai plutôt tendu vers cette réconciliation entre les règles arabes et les signes étrangers, afin d'unifier le travail, d'en réduire le coût et d'en faciliter la tâche, d'autant plus que ces signes ont été largement utilisés dans les écoles, les publications et les manuscrits des pays arabes à notre époque.
[…]
La raison la plus importante pour laquelle on s'appuie sur ces signes est que les élèves égyptiens dans toutes les écoles publiques [dites École al-amiri pendant l'époque de Mohamed Ali Pacha], privées et étrangères apprennent ces signes tout en apprenant des langues étrangères. Si j'avais choisi d'autres signes, ce travail aurait distrait les étudiants, en particulier ceux qui débutent dans les études. Il y a là quelque chose qu'il faut éviter. » (Traduction de l'auteur)

[19] « Pour toutes ces raisons, j'ai vu la nécessité de s'appuyer sur ces signes, après avoir modifié leur dessin, pour qu'ils puissent être écrits avec le script arabe : en tenant compte du mouvement de la main dans l'écriture, [à savoir] de droite à gauche. » (Traduction de l'auteur)

[20] La norme ISO 646 est la première norme ayant succédé au code ASCII. Elle se présente en plusieurs versions nationales pour les langues (ou groupes de langues) européennes.

[21] Ceci apparait clairement en comparant la version ISO 646 modifiée pour l'arabe avec la version originale ISO 646 et les versions dites variantes nationales qui en découlent.

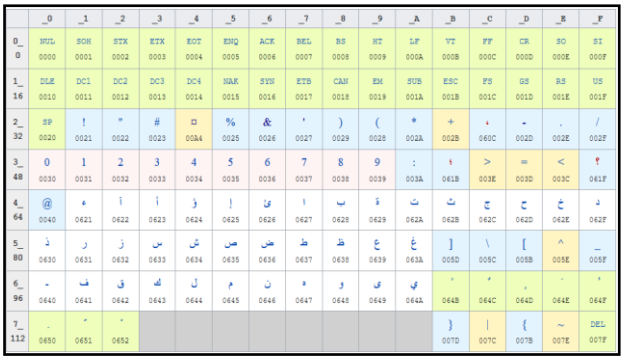
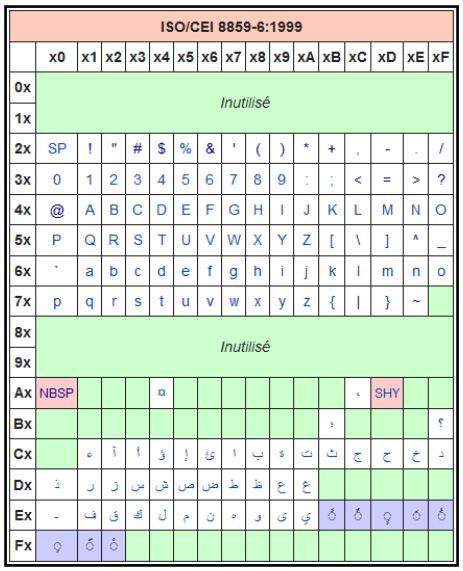


Figure 2.[22] ASMO 449 – Codage de 1982 (ISO 646 modifiée)

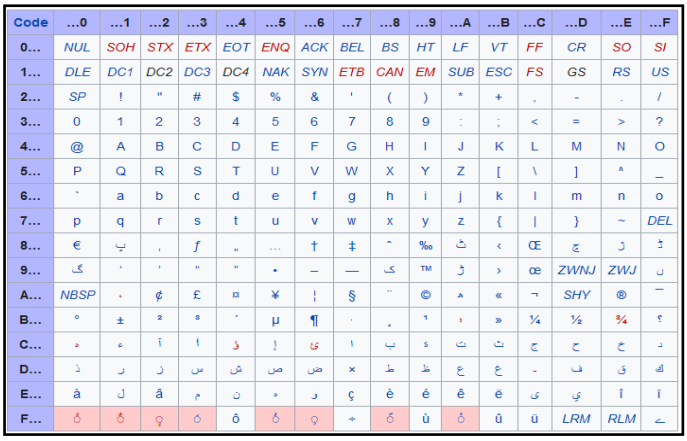
La question que nous soulevons est la suivante : si les signes de ponctuation sont universels, et utilisés dans la plupart des cas avec les mêmes glyphes, pourquoi alors ces trois signes de ponctuation arabes ne sont-ils pas traités comme tels au niveau du codage informatique ? En d'autres termes : pourquoi avoir créé, dans les codages postérieurs à l'ASMO 449 (voir Fig. 2 ci-dessus), de nouveaux caractères (donc de nouveaux codes) pour prendre en charge ces ponctuations, alors qu'il était possible de garder les mêmes caractères utilisés pour le latin et de laisser au logiciel (moteur de rendu) de choisir le glyphe adéquat selon la direction de l'écriture ? (Voir Figs. 3-5.)[23]



ISO/CEI 8859-6:1999															
	x0	x1	x2	x3	x4	x5	x6	x7	x8	x9	xA	xB	xC	xD	xE	xF
0x	*Inutilisé*															
1x																
2x	SP	!	"	#	$	%	&	'	(	)	*	+	،	-	.	/
3x	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	؟
4x	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5x	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6x	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7x	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	
8x	*Inutilisé*															
9x																
Ax	NBSP				¤								؛	SHY		
Bx												ء				؟
Cx		ء	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د
Dx	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ					
Ex	ـ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	ً	ٌ	ٍ	َ	ُ
Fx	ِ	ّ	ْ													

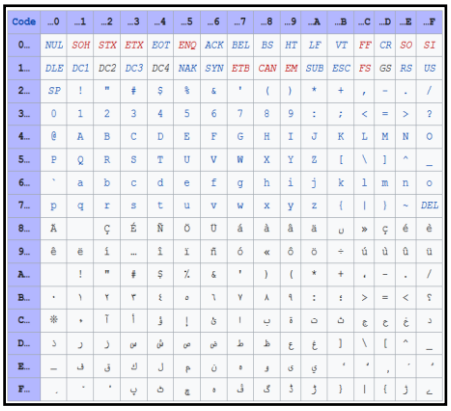
Figure 3.[24] Codage ISO/IEC 8859-6:1999 – langue arabe (table de caractères 8-bits – extension ISO 646)

[22] Source : http://www.langbox.com/arabic/asmo449.pdf
[23] Le codage MacArabic, utilisé dans les ordinateurs Apple Macintosh, est un cas extrême où tous les signes ponctuo-typographiques ont été recodés dans la partie haute réservée au script arabe.
[24] Source : https://www.ascii-code.com/ISO-8859-6



Code	…0	…1	…2	…3	…4	…5	…6	…7	…8	…9	…A	…B	…C	…D	…E	…F
0…	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1…	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2…	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3…	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4…	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5…	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6…	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7…	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL
8…	€	پ	‚	ƒ	„	…	†	‡	ˆ	‰	ٹ	‹	Œ	چ	ڈ	ڈ
9…	گ	'	'	"	"	•	–	—	ک	™	ڑ	›	œ	ZWNJ	ZWJ	ں
A…	NBSP	،	¢	£	¤	¥	¦	§	¨	©	ھ	«	¬	SHY	®	¯
B…	°	±	²	³	´	µ	¶	·	¸	¹	؛	»	¼	½	¾	؟
C…	ہ	ء	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د
D…	ذ	ر	ز	س	ش	ص	ض	×	ط	ظ	ع	غ	ـ	ف	ق	ك
E…	à	ل	â	م	ن	ه	و	ç	è	é	ê	ë	ى	ي	î	ï
F…	ً	ٌ	ٍ	َ	ô	ُ	ِ	÷	ّ	ù	ْ	û	ü	LRM	RLM	ے

Figure 4.[25] Page de code 1256 – script arabe (table 8-bits – Microsoft Windows)



Code	…0	…1	…2	…3	…4	…5	…6	…7	…8	…9	…A	…B	…C	…D	…E	…F
0…	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1…	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2…	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3…	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4…	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5…	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6…	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7…	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL
8…	Ä		Ç	É	Ñ	Ö	Ü	á	à	â	ä	ں	»	ç	é	è
9…	ê	ë	í	…	î	ï	ñ	ó	«	ô	ö	÷	ú	ù	û	ü
A…		!	"	#	$	٪	&	'	)	(	*	+	،	-	.	/
B…	٠	١	٢	٣	٤	٥	٦	٧	٨	٩	:	؛	>	=	<	؟
C…	❀	ء	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د
D…	ذ	ر	ز	س	ش	ص	ض	ط	ظ	ع	غ	]	\	[	^	_
E…	ـ	ف	ق	ك	ل	م	ن	ه	و	ى	ي	ً	ٌ	ٍ	َ	ُ
F…	ِ	ّ	ْ	پ	ٹ	ۃ	ء	ڤ	ڪ	ڈ	ڑ	)	|	{	ڑ	ے

Figure 5.[26] Page de code MacArabic – script arabe (table 8-bits – Mac OS)

Ce qui anime cette question est le fait qu'il existe une solution alternative qui a déjà été utilisée pour d'autres caractères et qui ne nécessite pas le recours à la création de nouveaux caractères (donc de nouveaux codes). Cette solution concerne tous les signes dont le glyphe peut être utilisé dans des textes bidirectionnels (de gauche-à-droite ou de droite-à-gauche) ou verticaux (de haut-en-bas ou de bas-en-haut) uniquement en lui faisant subir un effet miroir. L'un des principes fondamentaux d'Unicode est la distinction glyphe-caractère ; les caractères Unicode étant dotés de propriétés (qui les caractérisent) et d'algorithmes (qui permettent de les manipuler de manière appropriée). Pour montrer ces nouveaux concepts, prenons l'exemple de la parenthèse gauche qui a pour point de code U+0028. En pratique, elle est représentée par le glyphe « (» dans un texte latin (gauche-à-droite) et «) » dans un texte arabe (droite-à-gauche). C'est alors au moteur de rendu, muni du code du caractère en question (U+0028), de calculer la direction du texte et afficher le bon glyphe : « (» si le texte est de gauche à droite et «) » si le texte est de droite à gauche ; dans les deux cas, il s'agit bien d'une parenthèse ouvrante. Cette implémentation est réalisée à l'aide d'une propriété spéciale. Dans le standard Unicode, tous les caractères sont munis d'une propriété *Bidi_Mirroring_Glyph* qui indique si un caractère donné est à effet miroir ; c'est-à-dire s'il est utilisé dans les deux sens d'écriture uniquement en changeant de forme. Si tel est le cas, les moteurs de rendu utilisent

[25] Source : https://www.ascii-code.com/CP1256
[26] Source : https://www.wikiwand.com/en/MacArabic_encoding

un fichier appelé *BidiMirroring.txt* de la base de données des caractères Unicode pour récupérer l'œil miroir du caractère en question. (The Unicode Consortium, 2020 : p. 176-7) (Voir Tab. 1.)

Point de code	Glyphe	Nom du caractère	Œil miroir	Glyphe
0028	(	LEFT PARENTHESIS	)	0029
0029	)	RIGHT PARENTHESIS	(	0028
003C	<	LESS-THAN SIGN	>	003E
003E	>	GREATER-THAN SIGN	<	003C
005B	[	LEFT SQUARE BRACKET	]	005D
005D	]	RIGHT SQUARE BRACKET	[	005B
007B	{	LEFT CURLY BRACKET	}	007D
007D	}	RIGHT CURLY BRACKET	{	007B
00AB	«	LEFT-POINTING DOUBLE ANGLE QUOTATION MARK	»	00BB
00BB	»	RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK	«	00AB

Table 1.[27] Extrait du fichier *BidiMirroring.txt* – Partie 1 :[28] caractères appariés

En règle générale, un caractère ne doit pas avoir obligatoirement un œil miroir pour être doté de la propriété miroir. Ainsi, le fichier *BidiMirroring.txt* ne contient pas uniquement les caractères appariés, mais tout caractère ayant besoin de deux glyphes miroir. Cette exigence garantit que tout caractère ayant sa propriété miroir marquée (ayant comme valeur *vraie*) s'affichera correctement dans un contexte bidirectionnel. La figure 6 ci-dessous donne un aperçu du début de la deuxième partie du fichier *BidiMirroring.txt* qui contient tous les caractères non appariés pour lesquels les moteurs de rendu doivent fournir deux glyphes miroir pour assurer un affichage bidirectionnel.

```
FF62; FF63 # [BEST FIT] HALFWIDTH LEFT CORNER BRACKET
FF63; FF62 # [BEST FIT] HALFWIDTH RIGHT CORNER BRACKET

# The following characters have no appropriate mirroring character.
# For these characters it is up to the rendering system
#   to provide mirrored glyphs.

# 2140; DOUBLE-STRUCK N-ARY SUMMATION
# 2201; COMPLEMENT
# 2202; PARTIAL DIFFERENTIAL
# 2203; THERE EXISTS
# 2204; THERE DOES NOT EXIST
# 2211; N-ARY SUMMATION
# 2216; SET MINUS
# 221A; SQUARE ROOT
# 221B; CUBE ROOT
```

Figure 6.[29] Extrait de *BidiMirroring.txt* – Partie 2 : caractères non appariés

Le problème que nous venons de soulever ne concerne pas uniquement les trois ponctuations universelles. Tous les signes typographiques universels qui sont concernés par un ajustement selon la directionalité du texte sont concernés par le phénomène du double codage dans le standard Unicode et, par conséquent, d'une révision au niveau de la prise en charge logiciel.

Plusieurs avantages sont à tirer si les révisions du codage des graphèmes ponctuo-typographiques, comme nous les avons conçues ci-dessus, sont pris en charge dans les implémentations futures au niveau des modules de localisation des moteurs de rendu :

[27] Source : https://www.unicode.org/Public/10.0.0/ucd/BidiMirroring.txt
[28] Ce fichier est utilisé par les moteurs de rendu lorsque la valeur de *Bidi_Mirroring_Glyph* est égale à *vrai*.
[29] Source : Idem.

- Le fait de regrouper la majorité[30] des caractères de ponctuation communs aux différentes écritures dans un bloc séparé permet de faciliter la consultation des tableaux de caractères. Cette universalité et cette uniformité de codage permettent, en outre, une analyse, un tri, un affichage, un repérage et une édition efficaces des chaînes textuelles Unicode. (Andries, 2002 : p. 66)
- Cette conception favorise l'émergence d'une disposition universelle des caractères ponctuo-typographiques sur les claviers des ordinateurs.
- Cette solution préserve la distinction caractère-glyphe qui est l'un des dix principes fondamentaux qui ont guidés l'élaboration d'Unicode. (Andries, 2002 : p. 64)
- La concentration des caractères ponctuo-typographiques dans les deux blocs Latin (U+0000-U+007F) et Latin étendu (U+0080-U+008F), qui sont encodés avec un seul octet en UTF-8, permet une économie en matière d'espace de stockage (un octet au lieu de deux ou plusieurs).[31]

Références

- ANDRIES, P. (2002). « Introduction à Unicode et à l'ISO 10646 », Document numérique, 3(3-4), 51-88. Disponible sur [https://doi.org/10.3166/dn.6.3-4.51-88], consulté le 28/5/2021.
- THE UNICODE CONSORTIUM. (2020). « *The Unicode Standard, Version 13.0* – Core Specification », Mountain View, CA. ISBN 978-1-936213-26-9, Disponible sur [https://www.unicode.org/versions/Unicode13.0.0/], consulté le 28/5/2021.
- AḤMAD, Zakī. (2013). « Al-tarqīmu wa 'alāmātuhu fī al-luġati al-'arabiyyat », 1912, Le Caire : Kalimāt li-al-ṭarğamati wa al-našri.
- أحمد زكي. 1912. الترقيم وعلاماته في اللغة العربية، القاهرة: كلمات للترجمة والنشر، 2013.

Biographie de l'auteur

MAMMERI Mahmoud Fawzi est professeur à l'ESC Alger. Il a un doctorat en science en traitement automatique des langues. Il enseigne depuis plus de vingt ans l'informatique aux informaticiens et aux non informaticiens, notamment aux élèves de l'ESC Alger. Ses travaux de recherche portent essentiellement sur le TAL et la syntaxe de l'arabe. Il s'intéresse aussi au système d'écriture de l'arabe à l'ère du numérique. Il a été chercheur associé au CRSTDLA pendant une dizaine d'année et a participé à plusieurs projets de recherche dont un projet TASSILI de 2014 à 2017.

[30] Il est à noter que certains caractères de ponctuation sont spécifiques à certaines langues. Nous citons ici le cas de l'espagnol qui possède un système typique pour le marquage des formes interrogatives et exclamatives. En espagnol, une interrogation (exclamation) est introduite par un point d'interrogation (exclamation), de forme culbutée ¿ (¡), au début de l'interrogation (exclamation) et un autre, cette fois ci de forme normale ? (!), à la fin de l'interrogation (exclamation).

[31] UTF-8 (pour Format de transformation Unicode, en anglais *Unicode Transformation Format*) est l'un des encodages utilisés pour convertir les codes Unicode en unités (octets) stockables en mémoire.